

PLEASE RETURN TO  
MFC BRANCH LIBRARY

Technical Memo

INL Technical Library



138826

ANL/EES-TM-355, Vol. 1

ERROR PATTERN EVALUATION AND UNCERTAINTY  
QUANTIFICATION FOR A REGIONAL-SCALE  
(LAGRANGIAN STATISTICAL TRAJECTORY)  
ATMOSPHERIC TRANSPORT AND ACID-DEPOSITION MODEL  
Volume 1: Main Report

PROPERTY OF  
ANL-W Technical Library

RETURN TO REFERENCE FILE  
TECHNICAL PUBLICATIONS  
DEPARTMENT



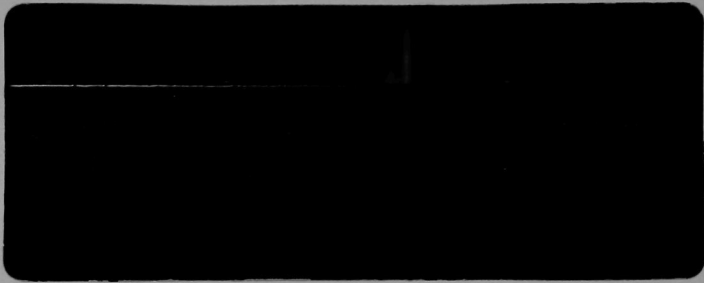
ARGONNE NATIONAL LABORATORY

Energy and Environmental Systems Division

Operated by

THE UNIVERSITY OF CHICAGO for U. S. DEPARTMENT OF ENERGY

under Contract W-31-109-Eng-38



Argonne National Laboratory, with facilities in the states of Illinois and Idaho, is owned by the United States government, and operated by The University of Chicago under the provisions of a contract with the Department of Energy.

#### **DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

This informal report presents preliminary results of ongoing work or work that is more limited in scope and depth than that described in formal reports issued by the Energy and Environmental Systems Division.



ARGONNE NATIONAL LABORATORY  
9700 South Cass Avenue, Argonne, Illinois 60439

---

ANL/EES-TM-355, Vol. 1

---

ERROR PATTERN EVALUATION AND UNCERTAINTY  
QUANTIFICATION FOR A REGIONAL-SCALE  
(LAGRANGIAN STATISTICAL TRAJECTORY)  
ATMOSPHERIC TRANSPORT AND ACID-DEPOSITION MODEL  
Volume 1: Main Report

by

Michael A. Lazaro and Jack D. Shannon\*

Energy and Environmental Systems Division  
Environmental, Technology, and Resource Assessment Programs

February 1988

work sponsored by

U.S. DEPARTMENT OF ENERGY and  
U.S. ENVIRONMENTAL PROTECTION AGENCY  
National Acid Precipitation Assessment Program

---

\*Biological, Environmental and Medical Research Division, ANL



## CONTENTS

### VOLUME 1: MAIN REPORT

FOREWORD .....	ix
ACKNOWLEDGMENTS .....	x
NOMENCLATURE .....	xi
SUMMARY .....	1
1 INTRODUCTION .....	9
2 REGIONAL ATMOSPHERIC TRANSMISSION, DEPOSITION, AND AIR CONCENTRATION MODEL .....	14
2.1 Brief Description of ASTRAP .....	14
2.2 Improvement in ASTRAP Model and Model Data Bases .....	15
2.2.1 Processing of Source Emission Fields .....	16
2.2.2 Processing of Wind and Precipitation Fields .....	16
2.3.3 Model Empirical Parameterization and Theoretical Formulation .....	16
2.3 Model and Model Data-Base Limitation .....	17
2.4 Model Adjustments for Performance Sensitivity Evaluation and Source- Receptor Uncertainty Analysis .....	19
3 MODEL INPUT AND EVALUATION DATA BASES .....	22
3.1 Source Emissions Data .....	22
3.2 Meteorological Data .....	26
3.3 Field Observations of Predicted Variables .....	28
3.3.1 Monthly Mean Air Concentration and Monthly Wet-Deposition Data .....	29
3.3.2 Seasonal Wet-Deposition and Ionic Concentration Data .....	31
3.4 Background Adjustments .....	33
4 OBJECTIVES AND METHODS FOR MODEL PERFORMANCE EVALUATION ....	39
4.1 Selection of Performance Evaluation Measures and Methods .....	40
4.2 Description of Performance Evaluation Measures and Data Analysis Methodology .....	41
4.2.1 Pattern Recognition through Data Display Techniques .....	42
4.2.2 Descriptive Statistical Performance Measures and Indices .....	47
5 MODEL PERFORMANCE AND SENSITIVITY EVALUATION RESULTS .....	52
5.1 Residual and Scatter Error Patterns .....	53
5.1.1 Monthly Average Air Concentrations .....	53
5.1.2 Seasonal Fluxes in Wet Sulfate Deposition .....	60
5.2 Sensitivity in Model Error Patterns .....	72
5.2.1 Seasonal Air Concentrations and Monthly Fluxes in Wet Sulfate Deposition .....	75

## CONTENTS (Cont'd)

5.2.2	Seasonal Fluxes in Wet Sulfate Deposition .....	84
5.3	Error Decomposition and Spatial Error Pattern Analysis .....	95
5.3.1	Separation and Computation of Bias, Temporal, and Spatial Error Components .....	95
5.3.2	Spatial Air Concentration and Deposition Error Patterns .....	103
5.4	Potential Factors of Influence on Apparent Model Performance .....	115
6	CONCLUSIONS AND RECOMMENDATIONS .....	127
6.1	Summary of Major Findings .....	128
6.1.1	Residual and Scatter Error Patterns .....	129
6.1.2	Performance Sensitivity Patterns in Apparent Model Error .....	131
6.1.3	Spatial, Temporal, and Bias Error Patterns .....	134
6.1.4	Potential Factors of Influence on Apparent Model Performance .....	135
6.2	Recommendations .....	136
6.2.1	Empirical Bayesian Probability Methodology .....	136
6.2.2	Estimation of Error in Data Bases .....	138
6.2.3	Model Sensitivity to Model and Data-Base Uncertainty Perturbations .....	139
	REFERENCES .....	142

## VOLUME 2: APPENDIXES A-N

APPENDIX A: Project Conference Papers

APPENDIX B: Summary of Previous Regional-Scale Sulfur Deposition Model  
Evaluation Studies

APPENDIX C: Source Emissions Data

APPENDIX D: Air Concentration and Wet Chemistry Sampling Networks

APPENDIX E: Parametric Statistical Performance Measures and Indices

APPENDIX F: Frequency Histograms

APPENDIX G: Corrections for Seasalt and Precipitation

APPENDIX H: Sensitivity Plots and Tables

APPENDIX I: Spatial Analysis Results

APPENDIX J: Influence of Protocol Sampling Period on Model Performance

APPENDIX K: Summary of Calculation Procedures

APPENDIX L: General Description of Simple and Universal Kriging  
Contour Plotting Method

## CONTENTS (Cont'd)

**APPENDIX M:** Predictions and Observations of Monthly Average Air Concentrations and Seasonal Wet Deposition and Site-Specific Background Values

**APPENDIX N:** Considerations on Using Stochastic Meteorology for Long-Range Atmospheric Transport Models

## TABLES

S.1	Summary of Nondimensional Indices of ASTRAP Performance for Predictions of Monthly Average Sulfate and Sulfur Dioxide Air Concentrations and for Seasonal Wet Sulfate Deposition .....	4
1.1	Elements of Uncertainty in an Integrated Acid-Deposition Policy Analysis .....	12
2.1	Base-Case Average and Range of ASTRAP Parameterizations .....	20
3.1	Distribution of Model Evaluation Air-Concentration Monitoring Sites by Month .....	29
3.2	Wet Sulfate Sampling Data Completeness Screening Criteria .....	33
3.3	Number of Sites that Meet Screening Criteria .....	34
4.1	Graphical Statistics Pattern Recognition Techniques .....	42
4.2	Descriptive Statistical Model Performance Measures and Indices .....	43
5.1	Internal Model Parameter Adjustments Used for Model Performance Sensitivity Tests .....	74
5.2	ASTRAP Dry-Deposition Velocity and Transformation Rate Variations, Normal and Adjusted .....	75
5.3	Sensitivity in Model Mean Sulfate and Sulfur Dioxide Air Concentrations to Variations in ASTRAP Internal Parameters .....	80
5.4	Sensitivity Ranges in ASTRAP for Four Seasons of Simulations .....	82
5.5	Relative Mean Square Error in Sulfate and Sulfur Dioxide Air Concentration Predictions .....	83
5.6	ASTRAP Sulfate and Sulfur Dioxide Air Concentration Seasonal Performance and Systematic Error Reduction with Parameter Variation .....	85
5.7	Sensitivity in Model Mean Wet Sulfate Deposition to Variations in ASTRAP Internal Parameters .....	89

## TABLES (Cont'd)

5.8	Sensitivity Ranges in ASTRAP for Eight Seasons of Simulations .....	91
5.9	Relative Dimensionless Mean Square Error and Rank Score Index in Wet Sulfate Deposition Predictions .....	92
5.10	ASTRAP Seasonal Performance and Systematic Error Reduction Potential with Parameter Variation .....	94
5.11	Temporal, Bias, and Spatial Error in ASTRAP Predictions of 1980 and 1981 Wet Sulfate Deposition .....	99
5.12	Explained Variance in ASTRAP Predictions .....	102
5.13	Comparison of Observed and Predicted Spatial Patterns of Wet Sulfate Deposition -- Simple Kriging for 1980 and 1981 .....	113
5.14	Comparison of Model Performance across Regions .....	118
5.15	Protocol Sampling Period Influence on Model Performance .....	120
5.16	Seasonal Frequency of ASTRAP Factor-of-Two Overpredictions at Model Evaluation Sites as a Function of Sampling Protocol .....	121
5.17	Comparison of Model Performance Based on Spatial Aggregation .....	124
5.18	Comparison of Model Performance Based on Precipitation-Weighted Ionic Concentration Versus Mass Flux .....	126

## FIGURES

1.1	An Integrated Acid-Deposition Policy Analysis Framework .....	10
3.1	Source-Receptor Grid for ASTRAP Model Evaluation .....	24
3.2	Seasonal 1980 and 1981 Sulfur Emissions by Grid Region .....	25
3.3	Seasonal 1981 Sulfur Emissions by Grid Region and Release Height .....	27
3.4	Sulfate and Sulfur Dioxide Air Concentration and Wet Chemistry Monitoring Sites in 1978 .....	30
3.5	Wet Chemistry Monitoring Sites in 1980 and 1981 .....	32
3.6	Natural and Man-Made Sulfur Emission Densities .....	36
3.7	Global Trend Network Sites .....	37
3.8	Mean Annual Ionic Concentrations of Sulfate at Global Trend Network Sites .....	38

## FIGURES (Cont'd)

5.1	Frequency Histograms of Monthly Sulfate Air Concentration Residuals .....	54
5.2	Frequency Histograms of Monthly Sulfur Dioxide Air Concentration Residuals .....	55
5.3	Monthly Scatter Plots of ASTRAP Predictions Versus Field Observations of Sulfate Air Concentrations in 1978 .....	56
5.4	Monthly Scatter Plots of ASTRAP Predictions Versus Field Observations of Sulfur Dioxide Air Concentrations in 1978 .....	57
5.5	Time-Series Plots of Average Monthly Sulfate and Sulfur Dioxide Air Concentration Mean Observations and Predictions .....	59
5.6	Grid Region Time-Series Plots of Average Monthly Sulfate Air Concentration Observations and Predictions in Selected Subregions .....	61
5.7	Grid Region Time-Series Plots of Average Monthly Sulfur Dioxide Air Concentration Observations and Predictions in Selected Subregions .....	62
5.8	Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for 1980 .....	63
5.9	Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for 1981 .....	64
5.10	Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for Combined Seasons in 1980, 1981, and 1980/1981 .....	65
5.11	Seasonal Scatter Plots of ASTRAP Predictions Versus Field Observations of Wet Sulfate Deposition in 1980 .....	67
5.12	Seasonal Scatter Plots of ASTRAP Predictions Versus Field Observations of Wet Sulfate Deposition in 1981 .....	68
5.13	Time-Series Plots of Average Wet Sulfate Deposition Observations and Predictions .....	70
5.14	Time-Series Plots of Average Wet Sulfate Deposition Observations and Predictions in Selected Subregions .....	71
5.15	Normalized Bias-Scatter Error Sensitivity Patterns for Sulfur Dioxide Air Concentrations .....	76
5.16	Normalized Bias-Scatter Error Sensitivity Patterns for Sulfate Air Concentrations .....	77
5.17	Fractional Error of Sulfur Dioxide and Sulfate Air Concentrations for Unpaired 1978 Predictions and Observations .....	79

## FIGURES (Cont'd)

5.18	Wet-Deposition Sensitivity Clusters for Summer 1980 .....	86
5.19	Fractional Error Sensitivity for Winter and Summer, 1980 and 1981 .....	88
5.20	Simple Kriging Contours of Observed and Predicted Sulfate Air Concentrations for January and July, 1978 .....	106
5.21	Simple Kriging Contours of Observed and Predicted Sulfate Air Concentrations for April and October, 1978 .....	107
5.22	Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Winter and Summer, 1980 .....	108
5.23	Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Winter and Summer, 1981 .....	109
5.24	Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Spring and Autumn, 1980 .....	110
5.25	Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Spring and Autumn, 1981 .....	111
5.26	ISDME Model Evaluation Region .....	114
5.27	Areas of Significant Differences between ASTRAP Predictions and Observations of Wet Sulfate Deposition in 1980 .....	116
6.1	Generalized Framework for Bayesian Uncertainty Analysis .....	139



## FOREWORD

This report is the result of studies funded as part of the National Acid Deposition Assessment Program (NAPAP) by the U.S. Department of Energy and, in earlier stages, also by the U.S. Environmental Protection Agency. The work originated as part of the studies that were managed by Task Group I, Assessments, and were aimed at supporting studies of the atmospheric sciences and source/receptor relationships, part of NAPAP's 1985 Assessment. A principal objective of those studies was to understand the performance characteristics of available atmospheric transport and deposition models for use in predicting the environmental impacts of sulfur emissions and in the possible development of optimal emission-control strategies. In particular, quantitative estimates of the uncertainty associated with certain aspects of model predictions are important to estimate the uncertainties in impact analyses and to determine the confidence that one should place in proposed policies and emission-control strategies. Model evaluation studies have often provided only limited insight into the nature of performance errors. Furthermore, the results sometimes do not address the performance factors of most importance to those using models for policy and assessment applications. In practical applications, the principal question is not why a model misses the mark, but by how much and in what manner it tends to miss. However, the apparent error between model predictions and observed values may also result in part from limitations in the observations, which needs to be taken into account in assessing the model's true performance.

The present study was intended to evaluate a range of statistical performance measures and to develop and test several new methods applied intensively to one model; however, the objectives and scope were still limited. Determination of the uncertainty in model-derived source/receptor relationships did not appear to be feasible from a comparison of the model predictions with the limited number of observations of deposition currently available. At best, it may be possible to estimate or place bounds on the uncertainty in predictions of deposition and to understand the character of those uncertainties. The performance of a model in predicting deposition and air-quality changes in response to large emission changes, outside the bounds of experience with historical evaluated data, cannot be confidently determined by the present methods. New approaches appear to be needed to provide a better understanding of the character of model errors and to estimate the uncertainties in future predictions within a limited domain, based on a comparison of predictions with historical observations. We hope that the results reported on here will be useful and can be applied to other models, as well as refined by further application with more extensive data.

Richard H. Ball  
Office of Environmental Analysis

## ACKNOWLEDGMENTS

Several people have made contributions to this project and to the completion of this report. We are especially indebted to R.H. Ball at the U.S. Department of Energy, Office of Environmental Analysis, for his guidance, creativity, patience, and technical assistance throughout the course of this work and for his review of and comments on the manuscript and drafts. We are also indebted to M.J. Bragen at Argonne National Laboratory (ANL) for his assistance in developing the statistical software package and for the data processing and to B.M. Lesht at ANL for modifying the ASTRAP model parameterization for our sensitivity study, providing the model output, and reviewing this report. D.L. Sisterson's (ANL) review and suggestions on Sec. 5.4, Potential Factors of Influence on Apparent Model Performance, are also appreciated.

## NOMENCLATURE

ACM	Aerochem Metrics
ADS	Acid Deposition System
AES	Atmospheric Environment Service
AMS	American Meteorological Society
ANL	Argonne National Laboratory
ANOVA	Analysis of variance
APIOs-C	Acid Precipitation in Ontario Study-Cumulative Network
APIOs-D	Acid Precipitation in Ontario Study-Daily Network
APN	Air and Precipitation Monitoring Network
ASTRAP	Advanced Statistical Trajectory Regional Air Pollution
BER	Bias error ratio
CANSAP	Canadian Network for Sampling Precipitation
$COV_1$	Covariance
$COV_2$	Residual covariance
$CV_o$	Coefficient of variation of observations
$CV_p$	Coefficient of variation of predictions
$CV_r$	Coefficient of variation of residuals
CVR	Ratio of coefficient of variation of predictions to the coefficient of variation of observations
DAC	Deposition and air concentrations
DMSE	Dimensionless mean square error
EBA	Empirical Bayesian approach
EBCV	Explained bias-corrected variance

EBM	Empirical Bayesian model
ELA	Experimental Lakes Area
FABE	Fractional average bias error
FBSE	Fractional bias and scatter error
FE	Fractional error
FSE	Fractional scatter error
GMLE	Geometric mean log error
GMSLE	Geometric mean square logarithmic error
GSDLE	Geometric standard deviation log error
GTN	Global Trends Network
HASL	Health and Safety Laboratory
IOA	Index of agreement
ISDME	International Sulfur Deposition Model Evaluation
LMSTD	Langrangian meso- and synoptic-scale transmission and deposition
MABE	Mean absolute bias error
MAP3S	Multistate Atmospheric Power Production Pollution Study
MBE	Mean bias error
MER	Model evaluation region
MLE	Mean logarithmic error
MOI	Memorandum of Intent
MPE	Model performance evaluation
MRPP	Mean residual-prediction product
MRR	Mean residual ratio
MSE	Mean square error

$MSE_a$	Additive systematic error
$MSE_i$	Interdependent systematic error
$MSE_p$	Proportional systematic error
$MSE_s$	Mean square error systematic
$MSE_u$	Mean square error unsystematic
MSES	Systematic mean square error
MSLE	Mean square logarithmic error
MSLTE	Mean square logarithmic temporal error
MSLSE	Mean square logarithmic spatial error
MSO	Mean square observation
MSP	Mean square prediction
MSSE	Mean square spatial error
MSO	Mean square spatial observation
MSSP	Mean square spatial prediction
MSTE	Mean square temporal error
MSTO	Mean square temporal observation
MSTP	Mean square temporal prediction
NADP	National Atmospheric Deposition Program
NAPAP	National Acid Precipitation Assessment Program
NB	Normalized bias
NBSE	Normalized bias and scatter error
NE	Normalized error
NMBE	Normalized mean bias error
NMC	National Meteorological Center

NS	Normalized scatter
NSE	Normalized scatter error
NWP	Numerical weather prediction
$\bar{O}$	Mean observation
$\bar{P}$	Mean prediction
PE	Potential error
PNL	Pacific Northwest Laboratory
P/O	Prediction-observation ratio
PS	Parameter set
PWIC	Precipitation-weighted ionic concentration
$R_1$	Correlation coefficient
$R_1^2$	Coefficient of determination or explained variance
$R_2$	Residual correlation coefficient
RAPT	Regional Air Pollutant Transport
RDMSE	Relative DMSE
REMSE	Relative mean square error
RMR	Residual mean ratio
RMSE	Root mean square error
RPRS	Relative prediction to residual scatter
RSI	Rank score index
SER	Spatial error ratio
SERP	Systematic error reduction potential
SLE	Standard deviation log error
SRR	Source-receptor relationship

$SS_b$	Sum of squares between groups
$SS_g$	Sum of squares within groups
SSE	Sum of squares error
SSR	Sum of squares regression
SSTO	Total sum of squares of the observations
STER	Ratio of spatial to temporal error
SURE	EPRI's Sulfate Regional Experiment
SVE	Spatial variance explained
$T_r$	Transformation rate
TCM	Tetrachloromercurate
TER	Temporal error ratio
TVE	Temporal variance explained
UAPSP	Utility Acid Precipitation Study Program
UDDBC	Unified Deposition Data Base Committee
UM	University of Michigan
$V_d$	Dry-deposition velocity
VAR	Variance
VLE	Variance logarithmic error
WC	Wet-removal coefficient
$\sigma$	Standard deviation
$\sigma_o$	Standard deviation of variations in observations
$\sigma_p$	Standard deviation of variations in predictions
$\sigma_r$	Standard deviation of the mean residual





**ERROR PATTERN EVALUATION AND UNCERTAINTY  
QUANTIFICATION FOR A REGIONAL-SCALE  
(LAGRANGIAN STATISTICAL TRAJECTORY)  
ATMOSPHERIC TRANSPORT AND  
ACID-DEPOSITION MODEL**

by

Michael A. Lazaro and Jack D. Shannon

**SUMMARY**

Knowledge of the relationship between causes and effects plays a key role when decisions involving scientific understanding and public policy formulation (science-policy decisions) are made. For instance, to achieve a particular improvement in environmental quality, policymakers need to know what causes the problems in order to decide what actions to take. Unfortunately, uncertainty is inherent in any decision-making process involving environmental issues. It is inherent in estimates of the types, probability, and magnitude of these adverse effects themselves. It is also inherent in estimates of the effects of the policy proposed to solve these problems -- both the economic effects and the existing and future environmental effects (on ecosystems and humans). The uncertainty in a proposed plan of action results from unpredictability caused by the policymaker's lack of knowledge or information and the incompleteness or unreliability of the information itself. This uncertainty can never be eliminated in science-policy decisions. In certain circumstances, however, it can be reduced, and it can be systematically analyzed and quantified so that the policy makers will be informed when deciding which options to consider and act upon. Deciding when an action can be taken is an issue based more on socioeconomics and political readiness than on the degree of scientific certainty.

This general understanding of the relationships among causes, effects, and science-policy decisions is relevant to determination of future policies on acid deposition. Knowledge, understanding, and ability in five important areas are prerequisites for developing, evaluating, and selecting the most cost-effective acid-deposition control and mitigation options. The areas are (1) an understanding of the causes of acid deposition from pollutant source-atmospheric interactions (source-term), (2) an understanding of the processes involved in the transfer of pollutants in source-receptor interactions (atmospheric term), (3) a knowledge of the impacts of deposition from atmosphere-receptor interactions (receptor term), (4) the ability to assess risks and perform a cost-benefit analysis (decision analysis term), and (5) the ability to estimate error and uncertainty (uncertainty analysis term) in the model estimates of sources of pollutant releases (causes), pollutant transmission and deposition, ecosystem damages and human health impacts (effects), and costs of deposition damages and pollutant controls. This study addresses limited aspects of the second and fifth areas (atmospheric term and uncertainty analysis term) by developing an analytical framework to describe and quantify error and uncertainty in long-range transport model predictions. The framework is composed of two components, a methodology for empirical Bayesian uncertainty

analysis and a methodology for parametric statistical error pattern analysis. The details and results of the latter component are given here, and the types and patterns of error are described and quantified.

One of the major goals of the National Acid Precipitation Assessment Program (NAPAP) is to provide objective and accurate estimates of the current contributions of anthropogenic and natural sources of acidic deposition, and estimates of the expected changes to these contributions from modifications in emission source strengths (NAPAP 1987a). This goal requires two types of "source-receptor relationships" (SRRs) to be developed: (1) SRRs that *apportion* the contribution from source areas (i.e., industrial centers, geopolitical regions) to receptor areas of acidic deposition (i.e., sensitive watershed, geopolitical regions) and (2) SRRs that *forecast response* in acidic deposition to receptor areas as a result of modifications to the emissions. The goal of this study is not to provide *apportionment* or *forecast response* SRRs but to provide a framework for quantifying the error and uncertainty in model predictions that may be used in providing SRRs for policy-decision analysis. The framework is composed of two components. The primary component uses an empirical Bayesian approach to quantify uncertainty in SRRs in probabilistic terms. This approach provides a means to compute the probability of the outcome (success or failure) of a set of proposed actions (i.e., policy options for control or mitigation of acid deposition) based upon the computed uncertainty in predicted variables pertinent to judging the success or failure of technologically and economically viable policy options. The ability to quantify the uncertainty in both the *apportionment* and *forecast response* SRR is of key importance to informed decision making. The second component of the framework uses some newly developed approaches to error pattern and error decomposition analysis along with some more traditional approaches to elicit patterns of the apparent error in model predictions. These approaches are not only important in expressing how well a model is performing but can also help provide, with accompanying model evaluation data-base and methodology enhancements, a better understanding of why a model performs the way it does. This understanding is critical to identifying and correcting the weak links in the modeling process. Because the empirical Bayesian approach has not been fully developed and tested, only the results from the second component of the framework for quantifying error and uncertainty in model prediction are reported at this time.

The principal objective of this study is to develop a flexible methodology to evaluate model performance that would help in the understanding of the characteristics and magnitude of the apparent error in model predictions. Because the characteristics of apparent model error can be highly complex, with multiple and interdependent causes, we adopted an approach that encompasses a combination of several new and existing statistical performance measures. This approach recognizes that no single or narrow group of performance measures (e.g., traditional distributional statistics designed to measure bias, correlation, and variance) can be used exclusively to uncover all the important characteristics of model performance. Our intention is not necessarily to determine why, in a *diagnostic* sense, a model performs well or poorly, although if causes of poor model performance can be identified and confirmed, we hope to be able to communicate this information to model developers so that model components and data-base elements can be improved. Rather, our task is simply to provide informative measures of how well or poorly, in an *operational* sense, a model performs under

different observable conditions and constraints. Results from model applications can then be quantified in terms of expected level of error. This study should, therefore, be viewed mainly from an application or *operational* perspective on evaluating and comparing models and model sensitivity rather than from a *diagnostic* or research-oriented perspective on improving model performance. Such information, if presented in a form that decision makers can understand and use, can have important implications for policy formulation and decision making.

Our specific goal is to develop a better understanding of the performance characteristics and apparent error of a long-term regional transport and deposition model. We intend to determine how well the predictions of the Advanced Statistical Trajectory Regional Air Pollution (ASTRAP) model compare with corresponding observations. Another goal is to discern and quantify differences in spatial and temporal patterns in seasonal and monthly mean observations and predictions and to determine the bias and scatter in model predictions. With respect to temporal patterns, we are interested in (1) how well the relative magnitudes of the peak seasonal wet sulfate ( $\text{SO}_4^-$ ) deposition and the peak monthly mean  $\text{SO}_4^-$  and sulfur dioxide ( $\text{SO}_2$ ) air concentrations (DAC) are reproduced in time, (2) how performance differs in seasons of the same year (interseasonal performance), and (3) how performance differs in seasons of separate years (interannual performance). With respect to spatial patterns, we are interested in (1) the location and magnitude of maxima, and (2) the location, orientation, shape, and gradient in the DAC contours.

The ASTRAP model evaluation data base used in this study consisted of monthly average  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations for four months in 1978, and seasonal wet  $\text{SO}_4^-$  deposition over a two-year period beginning in December 1979 and ending in November 1981. Greater physical understanding of model performance could be gained if predictions of wet deposition and air concentrations were evaluated for the same periods. For example, simulated atmospheric concentrations might be too low because parameterized wet removal is too high, but if simulated wet deposition for the same period is also too low, then some other feature must be involved. Similar deductive reasoning is possible if dry-deposition observations are also available for the same period. Unfortunately, suitable observation data sets for wet deposition, regional air quality, and/or dry deposition did not coincide. Nevertheless, the model evaluation methodology employed, along with model evaluation data base, did provide some useful and pertinent findings about the performance of the ASTRAP model.

Table S.1 summarizes the level of ASTRAP's performance when simulating monthly average  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations and seasonal wet  $\text{SO}_4^-$  deposition. The scatter error and bias error in model predictions are indicated by five nondimensional performance measures. The rank score error (RSI) is a performance measure designed to combine and balance the bias- and scatter-measuring attributes of the other four measures. The mean log error shows a strong tendency toward model overprediction of mean monthly  $\text{SO}_4^-$  concentrations in October and a lesser degree of model underprediction of mean monthly  $\text{SO}_2$  concentrations in January and April. The bias error is relatively small for the remaining monthly simulations. The same measures indicate a relatively large model overprediction of 1980 autumn and spring wet  $\text{SO}_4^-$  deposition and a lesser degree of model overprediction in the winter. The bias error is relatively small in the 1981 summer and spring. Overall the model performs best, as

**TABLE S.1 Summary of Nondimensional Indices of ASTRAP Performance for Predictions of Monthly Average Sulfate and Sulfur Dioxide Air Concentrations and for Seasonal Wet Sulfate Deposition**

Predicted Variable	Performance Index <sup>a</sup>				
	IOA	VLE	DMSE	MLE	RSI
<i>Air Concentrations</i>					
January 1978					
SO <sub>4</sub> <sup>=</sup>	0.60	0.229	0.132	0.105	2.13
SO <sub>2</sub>	0.62	0.232	0.309	0.254	2.41
April 1978					
SO <sub>4</sub> <sup>=</sup>	0.56	0.058	0.690	-0.118	2.65
SO <sub>2</sub>	0.61	0.163	0.294	0.240	2.35
July 1978					
SO <sub>4</sub> <sup>=</sup>	0.83	0.056	0.044	0.059	1.36
SO <sub>2</sub>	0.75	0.198	0.254	0.109	1.90
October 1978					
SO <sub>4</sub> <sup>=</sup>	0.45	0.076	0.251	-0.429	2.97
SO <sub>2</sub>	0.72	0.158	0.225	0.019	1.80
<i>Wet SO<sub>4</sub><sup>=</sup> Deposition</i>					
Winter					
1980	0.49	0.262	0.274	-0.246	2.82
1981	0.67	0.242	0.272	-0.197	2.20
Spring					
1980	0.63	0.121	0.242	-0.351	2.30
1981	0.82	0.116	0.114	0.015	1.47
Summer					
1980	0.85	0.138	0.161	-0.055	1.53
1981	0.85	0.100	0.110	0.054	1.44
Autumn					
1980	0.57	0.215	0.398	-0.405	2.79
1981	0.70	0.135	0.230	-0.315	2.11

<sup>a</sup>IOA = index of agreement (range 0.0. to 1.0; 1.0 = perfect model; 0.65 to 0.75 = average performance)

VLE = variance logarithmic error (performance improves as it approaches zero)

DMSE = dimensionless mean square error (range 0.0 to -5.0; 0.0 = perfect model)

MLE = mean logarithmic error (performance improves as it approaches zero)

RSI = rank score index (range 1.0 to -10.0; 1.0 = perfect model; 1.7 to 2.0 = average performance)

measured by RSI, in simulating July  $\text{SO}_4^-$  air concentrations. Performance on a comparatively high level is also exhibited for simulation of wet  $\text{SO}_4^-$  deposition for the summer and spring of 1981. Relatively poor performance is shown for October  $\text{SO}_4^-$  and January  $\text{SO}_2$  air concentrations and winter and autumn 1980 wet  $\text{SO}_4^-$  deposition.

The magnitude of maximum seasonal wet  $\text{SO}_4^-$  deposition and monthly mean air concentrations can be reproduced reasonably well. Without additional years of model evaluation data, however, it is not possible to determine whether temporal patterns in observed wet  $\text{SO}_4^-$  deposition are also reasonably reproduced by the model. The limited data analyses show significant interseasonal and interannual (between winters, springs, and autumns) differences in model performance. More data are needed to confirm this.

Perturbations of four model parameters within the estimated range of uncertainty of these parameters revealed that simulations of  $\text{SO}_4^-$  air concentrations are most sensitive to variations in transformation rate ( $T_r$ ), while  $\text{SO}_2$  concentrations are most sensitive to variations in dry-deposition velocity ( $V_d$ ). Wet  $\text{SO}_4^-$  deposition is most sensitive to variations in the wet-removal coefficient (WC). The systematic error reduction potential (SERP), through adjustments in these parameters, is most significant for simulations of October and January  $\text{SO}_4^-$  (SERP = 59% and 23%) and April  $\text{SO}_2$  (SERP = 23%) air concentrations and for simulations of autumn (41% in 1981) and spring (22% in 1980) wet  $\text{SO}_4^-$  deposition. This suggests that a significant fraction of the systematic error in model predictions for these monthly and seasonal periods, most notably October 1978 and autumn 1981, can be reduced through adjustments in model parameterization. Systematic error in model predictions may be associated not only with model parameterizations but also with the estimation of source emissions and the analysis of wind and precipitation fields. Errors or unrepresentativeness in verification data can lead to an apparent systematic error in model predictions. Without a means to segregate the sources of systematic error in model predictions, any revisions to model parameterization should be made cautiously and should be based upon relevant field measurement, of key processes. In Sec. 6.2, discussions stress the importance of quantifying model input and model evaluation data (field measurements of DAC) error, and recommendations are given for the statistical treatment of this error to more readily identify its sources (i.e., model parameterization, field measurement).

Error patterns are examined by decomposition of mean square error (MSE) into its spatial, temporal, and bias components and by decomposition of variance into spatial and temporal components. Kriging is then used to further examine ASTRAP's ability to reproduce spatial patterns in the observational fields (such as the position, shape, orientation, and magnitude of the gradient in the isolines). The spatial error in ASTRAP predictions of wet  $\text{SO}_4^-$  deposition dominates, accounting for over 70% of the total error in the winter, spring, and summer. The predicted wet  $\text{SO}_4^-$  deposition in the autumn shows comparative levels of spatial and bias error, with a relatively small contribution of temporal error to the total error. The temporal error across seasons is smaller than the other two error components, particularly for winter and autumn. These results are probably caused by the statistically small number of data points (two seasons) considered in our analysis. Although only 50% of the error in the autumn predictions is spatial in origin, the relative larger overall error in autumn (73% to 122% greater than in spring and summer) makes the spatial error in autumn slightly larger than that in spring and

summer. The computed bias-corrected variances (EBCVs) show that the model's ability to explain variance in summer simulations (over 40%) is substantially better than its ability to explain variance in other seasons. In fact, the negative EBCVs computed for the winter, spring, and autumn simulations show the model does not do well in explaining observed interannual variance for these seasons. This fact seems to indicate that the interannual correspondence between predictions and observations for nonsummer simulations are nearly random. These results should be viewed with caution because of the limited data available at the start of the study.

Results from the kriging analysis provided additional indications of ASTRAP's limitations in reproducing monthly and seasonal patterns in DAC, although the extent to which the various monitoring sites captured the regional patterns remains a contentious matter. This analysis showed that although the magnitudes of the observed maxima in  $\text{SO}_4^-$  air concentrations and wet  $\text{SO}_4^-$  deposition are reasonably reproduced, the locations of these maxima are not. The model had difficulty in reproducing the position, shape, and magnitude of the gradient in the observed spatial patterns. Although there are no significant variations in the magnitudes of the predicted and observed interannual maximum wet depositions, there are significant variations in the locations of the observed maxima. Since variability in meteorology plays an important role in influencing locations where observed maxima occur, the difficulty that ASTRAP and, indeed, all regional transport models have in properly characterizing local or subgrid variations in wind and precipitation fields may be an important contributing factor in the model's inability to accurately locate the maximum deposition areas. This difficulty in characterizing the stochastic nature of winds and precipitation may also play an important role in the poor reproduction of other spatial features in the observed data, such as the shape, orientation, and magnitude of the deposition gradient, although simplifications in parameterizations of chemical or removal processes may also contribute. This hypothesis needs to be investigated through an analysis of alternatives, such as numerical weather prediction models for generating mass-consistent and dynamically correct winds. Through use of a more complete and longer (through 1986) precipitation chemistry data base, many of these issues could be more readily addressed.

Our analysis of factors that may influence model performance indicates that sampling protocol can be a major contributor to the observed apparent model performance. Significant ASTRAP overpredictions (prediction-observation ratios greater than two) are more frequent for event or daily collectors than for weekly or monthly collectors, particularly during the colder seasons. This may be due, in part, to the more complete oxidation of S(IV) to S(VI) for collectors on longer sampling protocols. The cold temperatures and limited availability of oxidants in winter, with a resulting observed S(IV) maximum observed during this season (Dana 1980), would seem to explain the lower  $\text{SO}_4^-$  concentrations in samples that are preserved. The S(IV) would gradually be converted to S(VI) in samples that are not preserved. The wet-removal parameterization in ASTRAP is for bulk sulfur; i.e., removal rates for  $\text{SO}_2$  and  $\text{SO}_4^-$  are identical. The rationale for this is that while initial wet removal of  $\text{SO}_4^-$  in the atmosphere is more efficient than removal of  $\text{SO}_2$  (the  $\text{SO}_4^-$  aerosol serve as cloud condensation nuclei), in-cloud oxidation of  $\text{SO}_2$  can be rapid, especially in the summer when oxidants are plentiful. The wet sulfur deposition predicted by ASTRAP corresponds to the bulk sulfur equivalent of combined S(IV) and S(VI) and should, if S(IV) is not measured, more closely



correspond to observations in which sample preservation of S(IV) is not ensured (NADP, APIOS-C, CANSAP, etc., sampling networks).

When viewed from the perspective of providing *apportionment* and *forecast response* SRRs, the analysis, provided with the two years of data examined in this study, seems to indicate that ASTRAP's limited ability to reproduce the spatial patterns of seasonal wet deposition could affect its usefulness. A fairly strong seasonal dependence on model performance is also indicated (best during summer or spring for wet deposition and summer for  $\text{SO}_4^{2-}$  air concentrations). Because of the episodic nature of wet deposition and because of the model's better performance over periods where the cumulative episodic and nonepisodic deposition constitute a significant fraction of the total annual deposition, applications that require development of SRRs (for use in strategies requiring yearly deposition amounts) are still feasible. Additional years of model evaluation data and spatial error analysis are needed to better determine this possibility.

Although this study helps to provide new insights into model performance evaluation (MPE) methods and better understanding of MPE results, we are still unable to specify the level of uncertainty in model predictions, and we still lack a fundamental understanding of why long-range transport models perform the way they do. Three areas of further research could help provide a means to quantify uncertainty and to improve our understanding of model performance: (1) the completion of the development and test application of the empirical Bayesian uncertainty quantification methodology, (2) the estimation of model input and model evaluation data errors and the explicit incorporation of these errors into measures of performance of model predictions, and (3) the investigation of the sensitivity in model performance by local and global variation of an expanded set of key model and model input variables.





## 1 INTRODUCTION

The relationship between causes and effects plays a key role in decisions involving scientific understanding and public policy formulation (science-policy decisions). For instance, to achieve a particular change in environmental quality, policy makers need to know what causes the problems in order to decide what actions to take. Unfortunately, however, uncertainty is inherent in any plan that is proposed to reduce adverse effects. It is inherent in estimates of the types, probability, and magnitude of these adverse effects themselves. It is also inherent in estimates of the effects of the policy proposed to solve these problems -- both the economic effects and the existing and future environmental effects (on ecosystems and humans). The uncertainty in a proposed plan of action results from unpredictability caused by the policy maker's lack of knowledge or information and the incompleteness or unreliability of the information itself. This uncertainty can never be eliminated in science-policy decisions. In certain circumstances, however, it can be reduced, and it can be systematically analyzed and quantified so that the policy makers will be informed when deciding which options to consider and act upon. Deciding when an action can be taken is an issue based more on socioeconomics and political readiness than on the degree of scientific certainty.

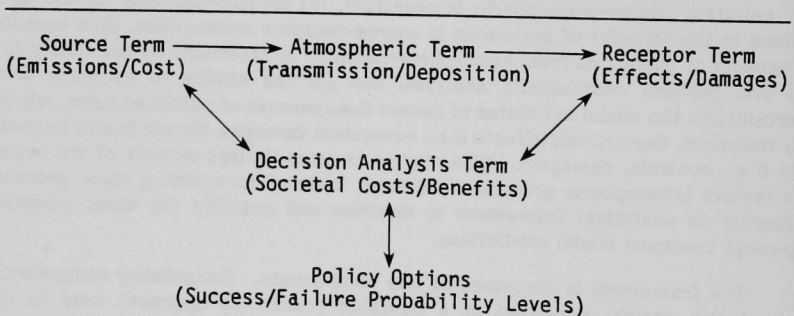
This general understanding of the relationships among causes, effects, and science-policy decisions is relevant and can be applied to future policies on acid deposition. Knowledge and ability in five important areas are prerequisites for developing, evaluating, and selecting the most cost-effective acid-deposition control and mitigation options. The areas are (1) an understanding of the causes of acid deposition from pollutant source-atmospheric interactions, (2) an understanding of the processes involved in the transfer of pollutants in source-receptor interactions, (3) a knowledge of the impacts of deposition from atmosphere-receptor interactions, (4) the ability to assess risks and perform cost-benefit analyses, and (5) the ability to estimate error and uncertainty in the model estimates of causes (i.e., sources of pollutant releases), transfer (i.e., transport, deposition), effects (i.e., ecosystem damages, human health impacts), and costs (i.e., controls, damages). This study addresses limited aspects of the second and fifth factors (atmospheric processes and error inherent to modeling these processes) by developing an analytical framework to describe and quantify the error uncertainty in long-range transport model predictions.

The framework is composed of two components. The primary component at the onset of this project consisted of a Bayesian probability approach used to quantify uncertainty in source-receptor relationships in probabilistic terms. Some of the preliminary Bayesian results were presented at a joint American Meteorological Society and Air Pollution Control Association Conference held at Chapel Hill, North Carolina, in November 1986 (see App. A). Because the Bayesian theory is still undergoing development and testing, the theory and more formal results of this approach are not discussed in this document. The second component uses some novel approaches along with more traditional parametric statistical measures of model performance and statistical graphics to elicit patterns of apparent error in source-receptor relationships. Some of the preliminary results of this component of our study were also presented at the conference (also in App. A). The details of this study are given in this report.

Several uncertain areas must be dealt with in any integrated acid-deposition policy analysis.\* These elements of uncertainty can be grouped into four "terms" (Fig. 1.1); each term is contingent on the others, and each affects the policy options that could be considered. Each term forms an independent system of models and data bases for policy analysis. Examples of the models and data that make up each term are discussed in the following paragraphs.

Primary and secondary pollutant emission fields are defined as the source term. The spatial and temporal patterns in these fields can be estimated as retrospective (under existing controls) or prospective (under alternative control scenarios) source emission configurations with a variety of source models. These models can also provide cost estimates for various control options under consideration.

The atmospheric component includes meteorological/climatological data and an atmospheric model to process this and other input data, including source data, in order to compute spatial and temporal pollutant patterns. Several regional-scale transport models exist; they are generally of three basic formulations: Eulerian, Lagrangian, or statistical. The Eulerian models are generally more sophisticated in their treatment of atmospheric chemistry and physics and are usually applicable to simulation of episodic events. The Lagrangian or statistical models generally employ linear first-order chemistry and treat atmospheric physics in a highly parameterized manner. These models are usually applicable to longer time periods, on the order of a month to a year. There are also hybrid combinations of the above three basic model formulations.



**FIGURE 1.1 An Integrated Acid-Deposition Policy Analysis Framework**

---

\*The complexity and uncertainty inherent in the many issues associated with acid deposition demand the development of an integrated methodology or framework to facilitate decision making. Details on integrated assessments for acid-deposition policy formulation can be found in Lazaro et al. (1986) and Marnicio et al. (1985).

The third component of the evaluation matrix is the receptor term. Ecological, materials, and health dose-response data (where available), along with output data from atmospheric models, may be used with receptor models to compute spatial and temporal effects on aquatic and terrestrial ecosystems, materials, and human health. These models, or the output from these models can, in principle, be used to compute economic damages resulting from the projected effects.

The last element considered is the decision analysis term. Decision analysis models use societal cost-benefit data along with the computational output from the previous terms, and where possible, use available observational data (e.g., precipitation chemistry measurements) to compute and assign probable error for uncertainty distributions in projected decision-related variables.

Each component of an integrated acid-deposition policy analysis has reducible and irreducible levels of uncertainty. Reducible uncertainty is defined as uncertainty that can be quantified in some statistical fashion and measurably reduced by improving model and model data bases through research and development (e.g., field data collection and analysis). Irreducible uncertainty is defined as the spatial, temporal, and economic limit for collection of the field data necessary to improve the understanding of source-receptor relations (these are interrelationships among the source, atmosphere, and receptor terms). This uncertainty is of a stochastic nature and prohibits the unique specification of the state of the atmosphere and biosphere in any given region of space and time. Table 1.1 identifies the specific elements of uncertainty in each of the terms.

This study deals primarily with the composite or aggregate error coming from the atmospheric component of an integrated assessment. The error (or difference between the predictions of a model and the actual deposition or air quality) results from the combined effects of errors or limitations in the input data (such as emissions and meteorological data, where relevant) and errors in the algorithms of the model that represent physical and chemical processes. The principal objective of model evaluation for assessment purposes is to estimate the probable magnitude and character of this type of error. The error in the source term is implicit in the source emissions input data to the atmospheric model. Attempts are made to segregate and characterize the error in the field observational data used to evaluate model performance. Proper interpretation of results is difficult because this error may not be sufficiently quantified and may not be fully segregated from model and data input error. Errors and limitations in field observational data used to evaluate the model's performance interfere with estimation of the model error, usually increasing the apparent error between the observations and the predictions. Hence, failure to adequately account for this observational error tends to increase estimates of model error.

The principal causes of uncertainty in modeling the regional transport and dispersion of atmospheric pollutants are (1) the inherent spatial and temporal variability in the planetary boundary layer, (2) the stochastic nature of the relationships between atmospheric state variables and the measurement of meteorological variables used to approximate atmospheric state, and (3) the general lack of sufficient understanding of relationships between atmospheric conditions and the physical and chemical mechanisms

TABLE 1.1 Elements of Uncertainty in an Integrated Acid-Deposition Policy Analysis

- I. Source Term - Treatment of primary pollutant mass flux to atmosphere, and the economic and societal cost incurred in modifying that mass flux

Reducible

1. Determination of source strength and temporal/spatial variation  $Q(x,t)$
2. Determination of source physics - dynamic operating parameters (e.g.,  $V_g [t]$ ,  $T_g [t]$ ) and static characteristics (e.g.,  $d_g$ ,  $h_g$ )
3. Determination of reduction in source strength and the allocation of that reduction for various control strategy options  $\Delta Q(x,t)$
4. Determination of control costs

Irreducible

Specification of the state of *anthroposphere* (primary human activity) -- the infinite amount of data that would be necessary to specify the state of the man-made and natural source emission field uniquely in any given region of space and time cannot be acquired

- II. Atmospheric Term - Treatment of transport, diffusion, transformation, and removal for the prediction of concentration and deposition fields

Reducible

1. Formulation and parameterization of model equations (e.g., the treatment of horizontal and vertical dispersion, dry deposition velocities, transformation rates, wet removal rates, vertical mass exchange rates, and depth of horizontal transport layer uses empirical data to reduce actual concentration/deposition fields to an analogues set of concentration/deposition fields with finite degrees of freedom)
2. Numerical technique selected for the solution of governing equations
3. Empirical and theoretical data - The accuracy of measurement methods, and the limitations in the spatial and temporal resolution of those measurements
  - Horizontal winds and precipitation fields
  - Vertical profiles of wind, temperature, and humidity
  - Cloud spatial characteristics and type
  - Vertical profiles of cloud water and hydrometeor concentrations

Irreducible

Specification of the state of the *atmosphere* -- the infinite amount of data that would be necessary to specify the state of the atmosphere uniquely in any given region of space and time cannot be acquired

- III. Receptor Term - Treatment of deposition, assimilation, transport, diffusion, and physical/chemical/biological interaction for the prediction of ecological and health response

Reducible

1. Formulation and parameterization of model equations - dose-response functions
2. Solution of governing equations
3. Empirical and theoretical data - limitations to spatial and temporal resolution and accuracy in measurement and prediction methods
  - Ambient and ionic concentration data and wet-deposition data used to evaluate atmospheric models and drive ecologic models
  - Ecological and health effects data used to evaluate model performance
4. Determination of economic benefits of control

Irreducible

Specification of the state of the *biosphere* -- the infinite amount of data that would be necessary to specify the state of the biosphere uniquely in any given region of space and time cannot be acquired

of transport, transformation, scavenging, and deposition. Equally important are inherent uncertainties in the spatial and temporal distribution of source emissions and in the observational data for validating model performance, because they are significant factors to consider when making policy decisions. Acid-deposition assessments that rely on the outputs from models will require the uncertainty in model predictions to be characterized and quantified in a form understandable to decision makers. A means to do just that would be extremely useful (along with other tools such as cost-benefit analyses, risk assessments, and evaluations of control impacts on fuel markets and on employment) in providing information necessary for policy formulation.

The principal objective of this study is to develop a flexible methodology to evaluate model performance that would be useful in helping us to understand the characteristics and magnitude of the apparent error in model predictions. Because the characteristics of apparent model error\* can be highly complex, with multiple causes, an approach that encompasses a large and varied set of statistical performance measures was adopted. This approach recognizes that no single or narrow group of performance measures (e.g., traditional distributional statistics designed to measure bias, correlation, and variance) can be used exclusively to uncover all the important characteristics of model performance. The goal is not necessarily to identify specific weak links in model or data-base elements so that the weaknesses can be systematically removed. Although it is true that the identification of weak links may help us better understand and interpret certain patterns in model performance, our goal is simply to provide informative measures of how well or poorly a model performs under different observable conditions and constraints. Then results from future model applications can be qualified in terms of expected level of error or uncertainty. Such information, if presented in a form that decision makers can understand and use, can have extremely important implications for policy formulation and decision making.

Although we have applied the above approach to evaluating the performance of an atmospheric long-range transport and deposition model, it can be adapted, with appropriate observational data bases, to source term and receptor term models.

Section 2 of this report describes and gives the basis for selecting the atmospheric model evaluated in this study. Revisions and improvements (made from 1982 through 1985) and limitations in the model and model data base are described. The internal model parameter adjustments made for the model performance sensitivity evaluation are also discussed in Sec. 2. The data bases used as model input and used to evaluate model output are described in Sec. 3. Section 4 provides the general philosophy of the model evaluation approach and the specific objectives for evaluating the performance of the model. The model evaluation methodology, which includes graphical pattern recognition techniques, distributional parametric statistics, and error decomposition, is described in detail. The model performance results are given in Sec. 5, with a summary of findings and recommendations provided in Sec. 6.

---

\*Apparent model error is the error in model predictions associated with model input data, model parameterization and formulation, and the model evaluation observational field; it can be seen when comparisons of model predictions with corresponding field observations are analyzed.

## 2 REGIONAL ATMOSPHERIC TRANSMISSION,\* DEPOSITION, AND AIR CONCENTRATION MODEL

The Advanced Statistical Trajectory Regional Air Pollution (ASTRAP) model was chosen for this study for a number of reasons. Among them is its capability for rapid, economical, and efficient computations. This attribute is especially desirable when model sensitivity computations are required. Also, earlier versions of ASTRAP were subjects of previous model evaluation and sensitivity studies. These provide a reference point to measure any model or data-base improvements. Finally, ASTRAP's easy accessibility (an in-house model) and its availability in the public domain played a major role in its selection.

### 2.1 BRIEF DESCRIPTION OF ASTRAP

The ASTRAP model (Shannon 1981) consists of three major subprograms, plus various preprocessors and postprocessors, generally specific to the model application. In one subprogram, simulated trajectories are calculated for a grid of initial locations covering North America with a National Meteorological Center (NMC) spacing (300-375 km). The trajectory subprogram requires time-series plots of transport-wind and precipitation fields, generally organized seasonally. Ensemble statistics are produced from trajectory sets from four releases daily at each initial location for the period of simulation. The mean position and spread of the ensemble trajectories are described by fitting bivariate normal puffs to the end point ensembles as a function of time since release (plume age). Similar horizontal distribution functions are calculated for occurrences of wet deposition.

In a second, independent subprogram, synthetic, horizontally uniform, and diurnally repeatable meteorological data and parameterizations are used in a one-dimensional, vertical integration in which linear chemical transformation, vertical diffusion, and dry deposition are treated. Separate sets of calculations are made for initial emissions within each of the six model layers in the bottom 800 m of the atmosphere. Simulated releases are made throughout the diurnal cycle, and the results are averaged. The statistics stored in the vertical integration subprogram, as a function of plume age and emission layer, include one-dimensional concentrations of primary and secondary pollutants, total airborne pollutant burden, and dry deposition increments.

Since both the trajectory and vertical integration subprograms produce statistics for normalized or unit emissions, they must be combined with an emission inventory as well as with each other in order to calculate concentration or deposition values. This is accomplished in a third subprogram. The concentration and deposition subprogram, for each horizontal element in an emission inventory, selects the trajectory statistics from the closest initial location in the trajectory subprogram, adds a bias to correct for any difference between the source location and the trajectory initial location, and combines

---

\*Transmission = transport and diffusion, chemical transformation, and wet and dry removal (precipitation scavenging, dry-deposition physics, etc.).



them with the one-dimensional statistics for the appropriate emission layer. For concentrations, the two-dimensional puffs are combined with the one-dimensional surface concentrations of primary and secondary pollutants. For dry deposition, the two-dimensional puffs are combined with the one-dimensional dry-deposition increments for primary and secondary pollutants and converted to equivalent sulfur or nitrogen, as appropriate. For wet deposition, the two-dimensional puffs are combined with the one-dimensional airborne budgets, already converted to equivalent sulfur or nitrogen. The two-dimensional puffs are scaled by the emission rate.

The atmospheric concentrations predicted by the ASTRAP model correspond to an arithmetic average over the period of simulation, not a geometric average. While the nature of the continuous bivariate normal density functions fitted to trajectory end-point or wet-deposition ensembles is such that predicted values change even if the location shifts only slightly (provided enough significant digits are used in calculations), the real spatial resolution of ASTRAP simulations of concentrations and deposition is a complicated function of the resolutions of the emission inventory, the wind field, and the precipitation field. While the resolution of the wind field is coarsest (grid spacing 300-375 km), wind variability accounts for only a portion of the variance in simulations. Although this is only argued heuristically, we assume that the resolution of the ASTRAP model is somewhere between that of the wind field and that of the precipitation and emissions fields (typically 100-125 km). For simulations with exactly located point sources rather than an emissions grid, resolution would be somewhat better.

The ASTRAP model predicts deposition rather than concentration in precipitation, although precipitation-weighted concentration can be estimated by dividing the deposition values by the precipitation totals for the period of simulation. If the precipitation totals are obtained from the time series of precipitation analyses used in model calculations, the method is mass consistent. If the precipitation totals are taken from observations at monitoring sites, to the extent that the point concentration is used to imply an areal average, some mass inconsistencies may arise.

## 2.2 IMPROVEMENT IN ASTRAP MODEL AND MODEL DATA BASES

It would be useful to look at the evolution of the ASTRAP model over the past several years to identify improvements in model components. A convenient reference point from which to do this is the model application and performance evaluation study conducted under the U.S./Canada Memorandum of Intent (MOI) on Transboundary Air Pollution (Schiermeier and Misra 1982).

Model evaluation results for eight regional-scale sulfur transport and deposition models developed by U.S. and Canadian scientists were reported under the MOI study in November 1982. The study used a standardized set of 1978 emissions and meteorological data to compute monthly (January and July) ambient sulfate ( $\text{SO}_4^{2-}$ ) and sulfur dioxide ( $\text{SO}_2$ ) concentrations at 54 SURE network sites (Electric Power Research Institute's Sulfate Regional Experiment) and wet sulfur deposition at 3 U.S. MAP3S sites (Multistate Atmospheric Power Production Pollution Study) and 10 CANSAP sites (Canadian Network for Sampling Precipitation). The ASTRAP model was one of the eight models evaluated during the MOI work.

Significant improvements have been made to the data bases and to the empirical internal parameterization of ASTRAP since the MOI study. They are summarized here.

### **2.2.1 Processing of Source Emission Fields**

Emission estimates have changed periodically, particularly in the western states. In some of the early ASTRAP MOI work, the emissions for each state or province were represented by only a single virtual source. In this report, not only is a 100- to 125-km emission grid used in all cases, but also the virtual source for each cell is the emission-weighted centroid rather than the geometrical center as in the other MOI work, and the initial spread is calculated rather than assumed. These changes are probably most helpful when receptors are near major source regions, as Whiteface Mountain is near Montreal. In addition, the assumption of a distribution in the ambient wind when calculating effective plume height now causes the emission gridding to be much less sensitive to small changes in stack parameters than was previously the case.

### **2.2.2 Processing of Wind and Precipitation Fields**

Wind and precipitation analyses (January and July 1978) for the MOI work were performed by the Canadian Atmospheric Environment Service (AES). The analysis and preprocessing of meteorological data for this study were done by the University of Michigan (UM) and Argonne National Laboratory (ANL). This work is described in more detail in Sec. 3.2 of this report. The AES wind data provided for the MOI analysis were probably more reliable than the UM analysis data when applied to data-sparse areas, such as over the oceans, since the UM analysis was a by-product of a numerical weather analysis. A decision was made to modify ASTRAP to use the UM data instead of the AES data, since a longer period of record, eventually 24 years, was to become available. Only two full years of data (1980 and 1981) were available through AES. Although a detailed comparison of the AES and UM analysis methods was not made, it was felt that the quality obtained from both data bases was similar. However, no direct comparisons with ASTRAP were made to confirm that the two data sets produced comparable results, since the ASTRAP version compatible with the UM data incorporated several years of gradual model improvement.

### **2.2.3 Model Empirical Parameterization\* and Theoretical Formulation**

The basic structure of the ASTRAP model has remained the same since the MOI work. However, modifications have occurred to incorporate data in the ASTRAP parameterization schemes from more recent field experiments. These model modifications are briefly described here.

---

\*The term parameterization refers to a simplified, empirically based, functional form to represent physical and chemical atmospheric processes in the model.



1. *Dry-Deposition Parameterization.* Improved estimates of dry sulfate deposition were incorporated from the work of Wesely and Shannon (1984) and Wesely et al. (1985). The diurnal and seasonal patterns in sulfate deposition velocities are approximately 50% smaller than previous values. The deposition velocities for  $\text{SO}_2$  remain unchanged.
2. *Wet Removal Parameterization.* Modifications were made to coefficients to account for different methods of analyzing precipitation fields. The AES precipitation data had only 12-hour resolution; the wet removal algorithm assumed that all of that precipitation fell in one of the two 6-hour periods, and thus every other time step was, in effect, dry. No such assumption is necessary with the UM data. The convective transfer of pollutant mass to the free troposphere in precipitation processes was also adjusted from an assumed 20% venting to a 50% venting (i.e., 50% of the pollutant mass removed from the mixed layer was transferred to the free troposphere, where it was subject to subsequent wet removal but not dry removal). This adjustment seems reasonable in light of some recent data reported in the literature. Isaacs (1983) has computed a vertical transport ratio for various cloud types and frequency of occurrence over eastern Canada. From this formulation scheme, he estimates that on the average, 50% of sub-cloud air is pumped through its base each hour during the summer and 20% during the winter. However, no estimate on the fraction of vertical transport out of the mixed layer was given. Liu and McAfee (1984), looking at Ra-222 vertical distributions, concluded that 55% of continental Ra-222 during summer months and a much smaller amount during other seasons was transported out of the mixed layer.
3. *Trajectory and Vertical Integration Calculations.* ASTRAP now continues budget calculations to seven days after release rather than the previous five days.

## 2.3 MODEL AND MODEL DATA-BASE LIMITATION

Numerous factors contribute to errors or uncertainty in model predictions. These errors may originate from limitations in the input data used by the model, from limitations in the model's structural components, or from random, unpredictable fluctuations in the atmosphere.

Some of the key limitations contributing to uncertainty in the model and data base of this study are summarized here.

1. *Assumption of single-layer-averaged horizontal wind field.* Vertical averaging of wind velocity can induce computational errors in transport and dispersion. These errors are caused by an averaging of wind shear results, which can affect some layers more than others. The errors should be most significant to short-term ( $\leq 24$ -hour averages) predictions, and should be less important to long-term averages ( $\geq 3$ -month averages).
2. *Assumption of linear chemistry.* The assumption of first-order-linear kinetics may prove to be justified over seasonal periods, but aqueous-phase chemistry and convective cloud physics can have an important nonlinear influence on  $\text{SO}_2$  conversion chemistry over shorter time intervals (hours to several days). However, as the spatial and temporal averaging region increases, this and other nonlinear influences should play a less significant role.
3. *Oversimplification of wet removal processes.* Various parameterization schemes are used for modeling wet removal as a function of precipitation rate and a characteristic scavenging coefficient. Precipitation rates can be highly variable both spatially and temporally, especially during convective storms. Most model parameterization schemes do not account for the variation of a scavenging coefficient with (1) season, (2) size distribution of hydrometeors, (3) effective area of scavenged species, (4) snow type, and (5) storm type, partly for reasons of computational practicality and partly because many of these data are not routinely available over continental scales. These simplifications can distort spatial patterns in wet-deposition fields and force positive or negative bias in estimates of the magnitude of the deposition amounts.
4. *Oversimplification of dry-removal processes.* Dry-deposition schemes lack details on the spatial and temporal variations in deposition rates. (ASTRAP includes typical seasonal and diurnal variations in dry-deposition velocities but assumes these patterns are spatially uniform and repeated each day.) The influences of atmospheric stability and particle size are not explicitly accounted for in the parameterization schemes. Distortions in the spatial patterns and amounts of dry deposition, similar to those of wet deposition, can occur with these parameterization simplifications.
5. *Limited or nontreatment of vertical atmospheric motions.* Large convective storms can play a significant role in the redistribution of pollutants within the mixed layer and the venting of pollutants through the mixed layer. Uplift over frontal surfaces can elevate air initially in the mixed layer over wide regions. Not accounting for the redistribution from these and smaller systems can produce errors in (1) the horizontal pollutant transport, (2) the degree of

pollutant dilution, and (3) the amount of pollutants available for wet and dry removal. The ASTRAP model has a venting coefficient incorporated within its wet-removal parameterization but does not address the reentrainment of vented material back into the mixed layer.

6. *Omission of orographic effects.* Rough terrain and large water bodies can have a substantial influence on mesoscale meteorology and, therefore, on the transport, dilution, and removal of pollutants from the atmosphere. The wind and precipitation analysis techniques take orographic effects into account only implicitly and only to a small degree. Not accounting for terrain effects can produce errors in predicted spatial and temporal patterns of atmospheric concentrations and wet deposition.
7. *Spatial and temporal resolution of wind and precipitation measurements.* Wind data are derived from twice-daily upper-air observations over a sparse network (300- to 400-km spacing) of measurement stations. Precipitation observations use a denser network, but precipitation is also more variable. A lack of better meteorological resolution might have little effect on many long-term simulations, but errors related to geographic effects such as lake or seabreeze regions or precipitation gradients in mountainous areas will probably create bias.
8. *Accuracy of emission fields.* The nature of potential errors in preparing emission inventories is briefly discussed in Sec. 3.1.

## 2.4 MODEL ADJUSTMENTS FOR PERFORMANCE SENSITIVITY EVALUATION AND SOURCE-RECEPTOR UNCERTAINTY ANALYSIS

Except for highly statistical models that do not require temporally and spatially varying meteorological fields, model sensitivity and uncertainty studies in which all possible combinations of model parameterizations and modeling choices are tested are not computationally feasible. This is particularly evident when one considers that most parameterizations could span a continuous range rather than take only a few discrete values, and the sensitivity might vary from one year to the next. In these studies we have focused on four parameterizations, with the selections based upon commonality with the structure of other models, likelihood of eventual testing with field data, and general importance in deposition calculations. The parameterizations examined are the dry-deposition velocities for  $\text{SO}_2$ , dry-deposition velocities for  $\text{SO}_4^{2-}$ , linear rate of transformation of  $\text{SO}_2$  to  $\text{SO}_4^{2-}$ , and bulk wet sulfur removal coefficient. The dry-deposition velocity and transformation parameterizations are still given seasonal and diurnal patterns of variation, but the patterns are scaled to result in different average values. Similarly, the variation of the wet-removal coefficient during winter in northern latitudes is maintained, but the coefficient is scaled by the same factor used in other regions and seasons.

Varying the parameterizations tested was accomplished by scaling the seasonal and diurnal patterns and mean values by 0.5 and 2.0. Combinations of high, low, and base cases for the four parameterizations were tested, except for those that were deemed illogical (such as high  $\text{SO}_4^=$  dry deposition and low  $\text{SO}_2$  dry deposition, which in ASTRAP would result in a dry-deposition velocity for  $\text{SO}_2$  that was only half that for  $\text{SO}_4^=$ ). The average value and the extremes of the diurnal and seasonal patterns of the parameterizations whose sensitivities are studied in this report are given in Table 2.1.

**TABLE 2.1 Base-Case Average and Range of ASTRAP Parameterizations**

Parameter	Season	Average	High	Low
$\text{SO}_2$ dry-deposition velocity <sup>a</sup> (cm/s)	Winter	0.30	0.70	0.10
	Spring	0.40	0.80	0.10
	Summer	0.45	0.90	0.10
	Autumn	0.31	0.65	0.10
$\text{SO}_4^=$ dry-deposition velocity <sup>a</sup> (cm/s)	Winter	0.12	0.25	0.05
	Spring	0.20	0.40	0.05
	Summer	0.23	0.45	0.05
	Autumn	0.16	0.36	0.05
$\text{SO}_2$ to $\text{SO}_4^=$ transformation rate <sup>a</sup> (%/hr)	Winter	0.4	1.0	0.1
	Spring	1.2	3.0	0.3
	Summer	1.6	4.0	0.4
	Autumn	0.8	2.0	0.2
S wet-deposition coefficient (C) <sup>c</sup>	Winter	NA <sup>b</sup>	1.0	0.5
	Spring	1.0	1.0	1.0
	Summer	1.0	1.0	1.0
	Autumn	1.0	1.0	1.0

<sup>a</sup>These parameters vary diurnally in the ASTRAP model. The graphs describing these variations, along with their algorithms, can be found in Shannon (1985).

<sup>b</sup>NA = not applicable.

<sup>c</sup> $\text{CP}^{0.5}$   
Deposition = min  
0.50

P = precipitation (cm/6 hr).

In a linear model such as ASTRAP in which plumes and deposition patterns from multiple sources are superposed, the sensitivity in the modeled concentrations and deposition from multiple sources is relatively less than for typical single sources. For a fixed emission rate, any parameterization variation that reduces deposition in one area must increase deposition in some other area; thus, many perturbations in individual patterns "average out" when the patterns from many sources are superposed. The observations used to evaluate the uncertainties of model simulations combine the contributions from all sources, although some sources are more heavily affected by isolated major sources than are others. A large portion of model uncertainty must be associated with possible nonlinearities when there is an imbalance of  $\text{SO}_2$  and oxidizing species, but these nonlinearities cannot be examined in a linear model.

### 3 MODEL INPUT AND EVALUATION DATA BASES

Three major data bases were selected, analyzed, and screened for evaluating the performance of the ASTRAP model. These included  $\text{SO}_2$  emissions inventory, meteorological and ambient  $\text{SO}_2$  and  $\text{SO}_4^{=}$ , and wet sulfur precipitation chemistry data. The same data bases were used for the parametric statistical error analysis and the Bayesian uncertainty analysis of ASTRAP predictions. A description and analysis of these data bases follow.

#### 3.1 SOURCE EMISSIONS DATA

At the time the data-base needs were set for this project, Version 2.0 of the NAPAP emissions inventory was available. The NAPAP  $\text{SO}_2$  inventory consisted of data on both point and area source emissions over the continental United States. A point source was defined as any stationary source emitting at least 100 tons per year of any of the five primary-criteria pollutants with appropriate stack parameters needed to determine the effective point source height. These data were available on a stack-by-stack basis, and included coordinates, height, diameter, effluent temperature, flow rate, and annual emission rates. Seasonal emissions were computed from a monthly emissions inventory derived from the 1980 NAPAP inventory and source-category fuel-use patterns. All emission sources not fitting the point source definition were treated as area sources. These sources were primarily institutional, commercial, and residential fuel combustion sources; source industrial processes or space heating units were also included. Over 90% of the 1980 and 1981  $\text{SO}_2$  emissions in the United States came from point sources, with a majority of these emissions from the electric utility sector.

The Canadian  $\text{SO}_2$  emission inventory came from colleagues in the Canadian AES. Canadian  $\text{SO}_2$  emissions come primarily from metal smelters, which exhibit little regular seasonal dependencies. Approximately 1% to 2% of the annual U.S. and Canadian sulfur emission was assumed to be primary  $\text{SO}_4^{=}$ .

For many applications of the ASTRAP model, particularly those involving source-receptor matrices, it is important to have emissions spatially resolved more finely (i.e., to areas smaller than large geopolitical entities such as states or provinces). This is particularly true for predictions at monitoring sites or sensitive receptors lying within a source region (such as the Adirondacks in New York); unless the substate horizontal and vertical resolutions of emissions are known or estimable, all the emissions of a state or province are normally represented by only a single virtual source, which can lead to quite misleading near-source results. Knowledge of the horizontal source distribution for a distant upwind state (such as Missouri in the case of the Adirondacks) is perhaps of less importance, but even then good information on effective stack heights is vital because higher initial plumes result in less near-source dry deposition, thus leaving more pollutant remaining for downstream deposition, particularly in the wet form.

Temporal distributions of emissions, in particular seasonal patterns, are also important for modeling; overall sulfur oxides emissions in winter appear to be about 20% higher than for spring or fall. The overall secondary peak in  $\text{SO}_2$  emissions generally

occurs in the summer. However, a winter or summer peak in  $\text{SO}_2$  emissions will depend on the region (i.e., distribution of utility vs. nonutility source, fuel use) and on climate variations. The seasonal variations can be even larger for individual states, particularly those in which the combustion of fuel oil for home heating is important. The monthly emission inventories of Knudson (1985) provide excellent information on the seasonal variation of utility emissions (which in many cases do not vary greatly) but unfortunately little reliable information on the variation of miscellaneous sources, which would include residential space heating.

The emission preprocessor for ASTRAP grids the seasonal emissions from each state and province separately. The spatial resolution is the same in all cases; 100-125 km in the horizontal and 6 layers to 800 m in the vertical. It might appear that for a cell that overlapped geopolitical source regions (for example, the Kentucky-Ohio border), the Kentucky sources and the Ohio sources in the cell would have the same virtual source location. This is not the case, however. The mass-weighted emission centroid and Cartesian standard deviations are calculated for each state in turn; the overall emission grid thus has two virtual sources within that cell. For a medium-sized state such as Ohio, the horizontal distribution of emissions might be represented by 12 to 16 virtual sources. The vertical gridding is accomplished by using climatological fields of wind and temperature and a standard plume rise formula (Briggs 1971) to locate the mean effective height of the plume and then estimating an initial spread around that height by varying the wind speed by a factor of two. This procedure is done to account for the fact that the wind speed and consequent plume rise vary with actual meteorological conditions and is not an attempt to account for initial dispersion in the vertical, which is treated numerically in ASTRAP.

An emission inventory for 1980, roughly corresponding to NAPAP Version 2.0, was initially postprocessed by C. Benkovitz at Brookhaven National Laboratory (BNL) and then further processed to produce the emissions input for ASTRAP. For simulations specific to other time periods, scaling factors that related 1981 seasonal emissions (totalled statewide) to 1980 seasonal emissions and that related January, April, July, and October 1978 emissions to seasonal 1980 emissions (scaled to monthly averages) were developed from Knudson (1985). Some additional miscellaneous emission data sources for Canadian emissions were also applied for the SURE intensive periods. Except for some isolated Canadian point sources, all scaling factors were related to state or province totals and were applied to each source within the state or each source not individually treated in the province in the NAPAP-2.0/BNL/ANL gridded inventory.

The map in Fig 3.1 shows the source-receptor grid used for model performance evaluation. The eastern United States and southeastern Canada are divided into ten regions identified by Roman numerals in the figure. Each region is divided into four subregions of NMC dimensions (381 x 381 km at 60°N, 300-375 km over our latitudes of interest), and each subregion is divided into nine cells. Emissions were aggregated as appropriate. Tabulation of source emissions by season on a regional and subregional basis is given in App. C, Table C.1. Emissions outside the source-receptor grid area are not in Table C.1. However, these emissions (nongrid emissions) are shown in Fig. 3.2 for comparison with emissions from each of the ten regions. Regions II, III, and VI, the regions with the largest coal utilization, exhibit the largest seasonal variations. These



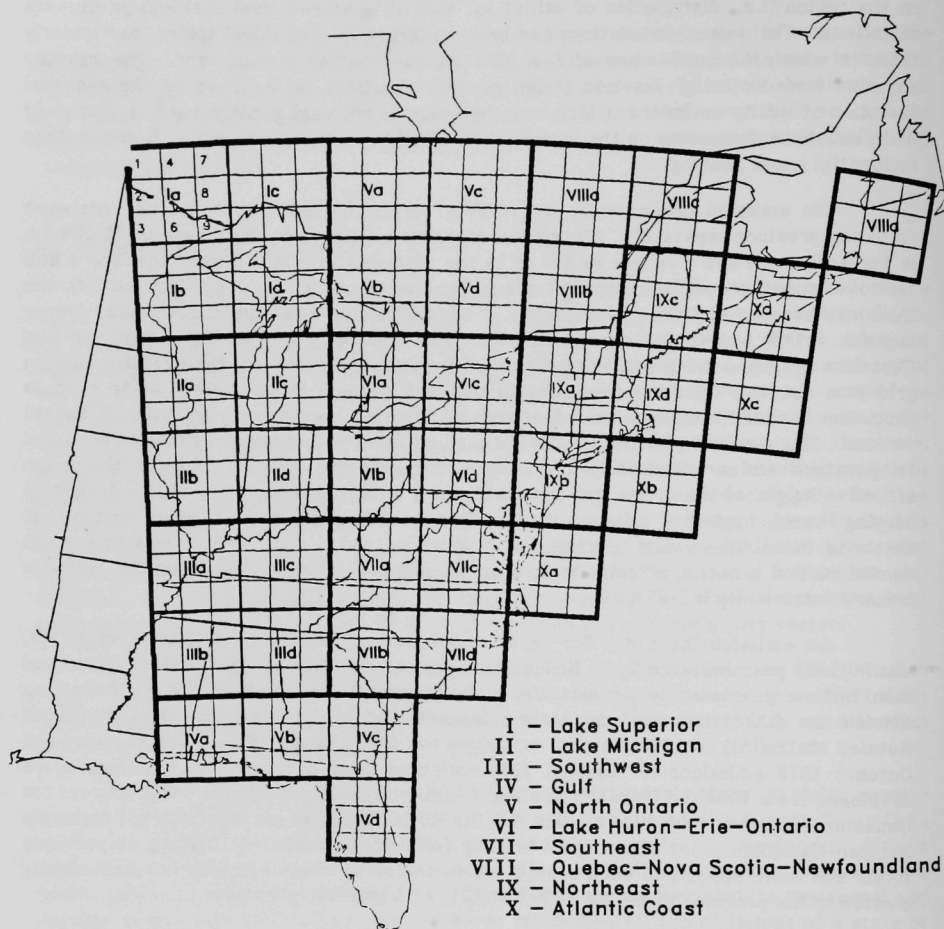


FIGURE 3.1 Source-Receptor Grid for ASTRAP Model Evaluation

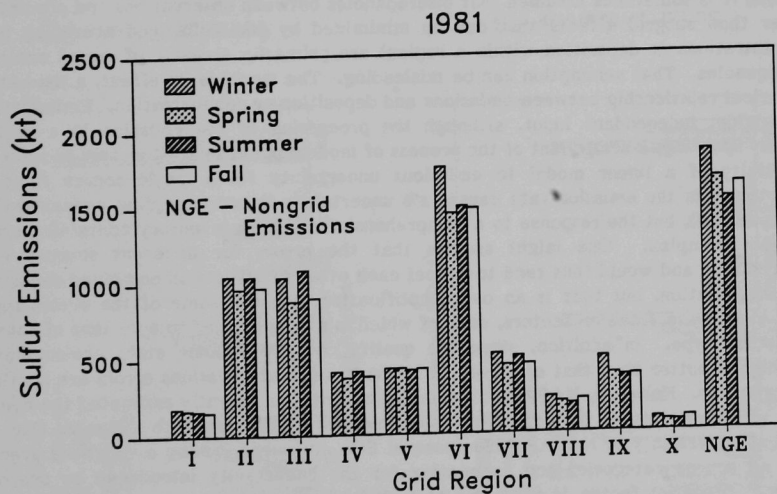
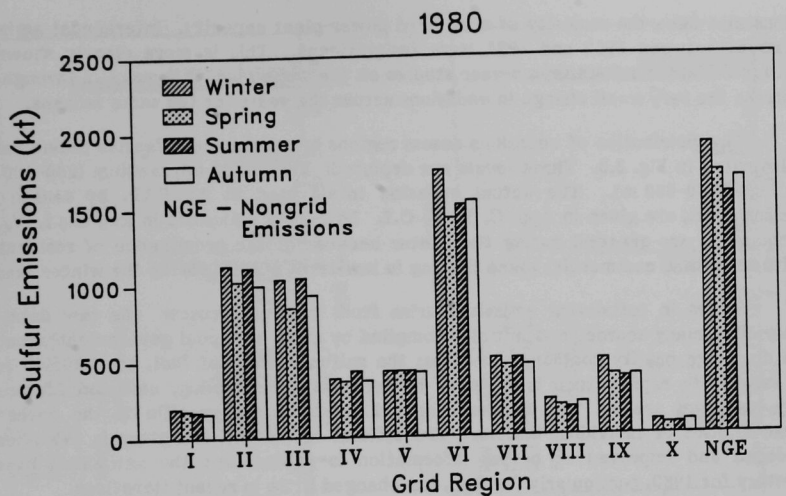


FIGURE 3.2 Seasonal 1980 and 1981 Sulfur Emissions by Grid Region

regions also have the majority of coal-fired power-plant capacity. Interannual emission variations between 1980 and 1981 were insignificant. This is more clearly shown by plotting like seasons for the two-year studies on the same plot. Figures C.2 through C.5 illustrate the very small change in emissions across the years for the same seasons.

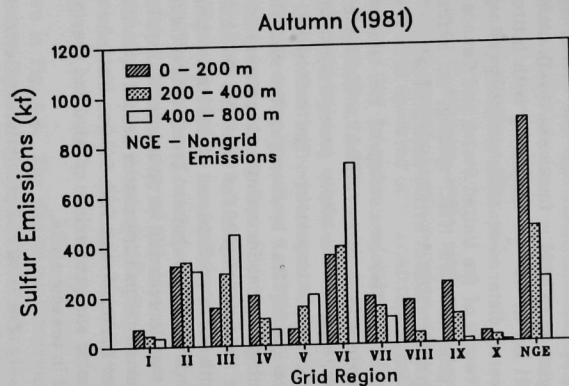
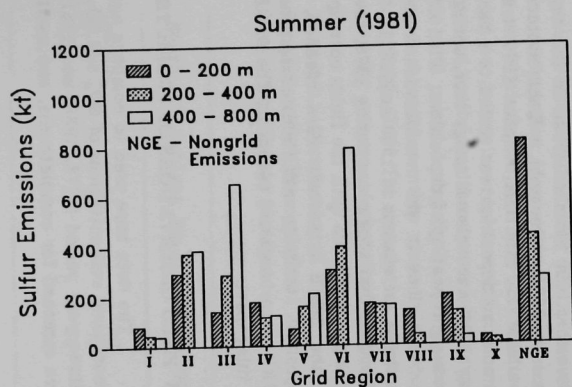
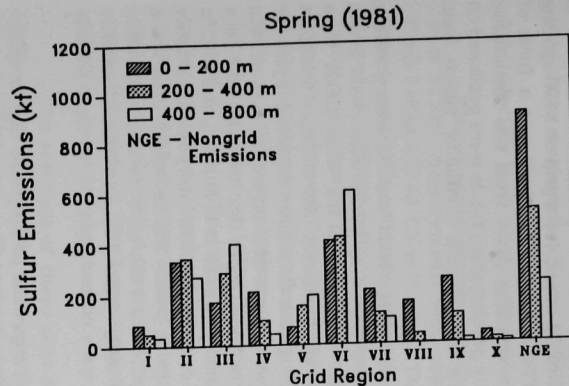
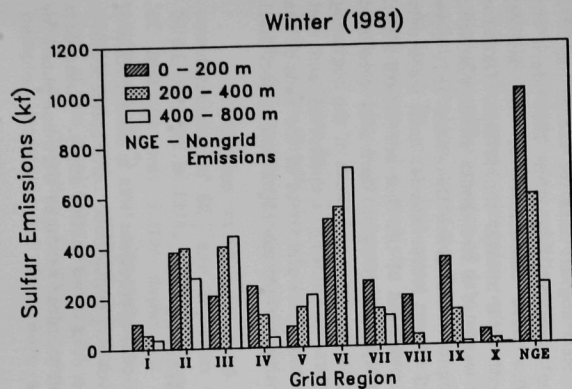
The distribution of emissions across regions by season and effective plume height is illustrated in Fig. 3.3. Three levels are depicted: low (0-200 m), medium (200-400 m), and high (400-800 m). The actual emission totals used in ASTRAP, by season and emission level, are given in App. C, Table C.2. Low-level emissions in Regions I, IV, and VII though X are greatest during the winter because of the prominence of residential, institutional, and commercial space heating (a low-level source) during the winter season.

Errors in estimating emissions arise from multiple causes: the raw data are gathered at many sources and initially compiled by state and local government agencies with disparate quality-control procedures; the sulfur content of fuel, particularly coal, can vary more rapidly than is revealed by intermittent sampling; emission factors in many instances are not well determined; and emissions are specific to the operating characteristics of individual boilers. The official or NAPAP inventory is periodically corrected and improved as better information is gained, but the estimated overall inventory for 1980, particularly for  $\text{SO}_x$ , has changed little in recent iterations.

In evaluations of regional sulfur atmospheric concentration and deposition models, it is sometimes assumed that discrepancies between observations and predictions (other than subgrid effects that can be minimized by combining and averaging point concentrations or deposition within a region) are primarily a result of model errors or inadequacies. That assumption can be misleading. The model is, in effect, a theoretical numerical relationship between emissions and deposition or concentration. Emissions are an outside, independent input, although the processing of the emission in a grid is, broadly speaking, a component of the process of model application. It is easy to show the sensitivity of a linear model to emissions uncertainty for a single source (i.e., x% uncertainty in the emission rate causes x% uncertainty in corresponding concentrations or deposition), but the response to a comprehensive emission inventory containing errors is more complex. One might assume that the errors for different sources were uncorrelated and would thus tend to cancel each other in effects on combined deposition or concentration, but that is an oversimplification, because some of the errors result from inadequate emission factors, each of which might be applied to a number of sources of similar type. In addition, since the quality control of some state environmental agencies is better than that of others, it is not likely that emissions errors are spatially homogeneous. However, NAPAP (interim assessment) has recently estimated the overall level of uncertainty in the source emission inventory. The estimate indicates that the level of uncertainty in NAPAP 1980 seasonal  $\text{SO}_2$  emissions, taking a weighted average over all source categories and accounting for the uncertainty introduced by use of a seasonal allocation factor, is  $\pm 13\%$  (NAPAP 1987a).

### 3.2 METEOROLOGICAL DATA

The meteorological data used in these ASTRAP simulations were initially processed under the guidance of P. Samson at the University of Michigan. There,



**FIGURE 3.3 Seasonal 1981 Sulfur Emissions by Grid Region and Release Height**

horizontal wind fields over the United States and Canada were produced for six 500-m layers up to 3000 m (mean sea level), at intervals of 12 hours, for a grid of NMC spacing (381 km at  $60^\circ$  N, increasing with latitude). The method of analysis was a form of inverse distance-squared weighting of upper air sounding data, from the rawinsonde networks of the United States and Canada. Hourly precipitation fields were produced on a grid with 1/3 NMC spacing, by averaging the observations (including zeros) from reporting stations within each grid cell.

Further processing of the wind and precipitation fields was necessary before their use as input to ASTRAP. The basic trajectory time step in ASTRAP is six hours; thus, sets of six hourly precipitation fields were totaled and the wind fields were temporally interpolated. The multi-layer winds were combined into a single transport layer for ASTRAP by averaging the wind fields through the first three reported layers for each grid point for spring and summer, the first two and one-half layers for autumn, and the first two layers for winter. The ASTRAP transport layer thus corresponded to a depth above ground level of 1500 m in spring and summer, 1250 m in autumn, and 1000 m in winter. The wind and precipitation fields were extrapolated and interpolated spatially so that there would be no missing values in rectangular grids. The technique used was a gradually expanding search around each missing value until one or more analyzed values were included, and then inverse distance-squared weighting. The locations of initially missing analyzed values of wind and precipitation were concentrated in northern Canada and in oceanic areas.

The wind and precipitation fields were then organized on magnetic tape in three-month data sets, with winter being December through February, spring being March through May, summer being June through August, and autumn being September through November. In the simulations in this report, meteorological data for the period 1976 through 1981 were used. Since meteorological data for December 1975 were not available at the time, winter 1976 trajectories were calculated for two months; cumulative deposition was scaled to a three-month total in later simulations. Data for May 1978 were also missing; thus, spring 1978 was also a two-month simulation with subsequent scaling of deposition. Wind and precipitation data for December 1981 were not used since it was considered a winter 1982 month (observational data were only available for seasons in 1980 and 1981). For simulations of the four one-month SURE intensives that took place during 1978, the quarterly meteorological data sets were used by skipping records until the fields corresponding to the dates and times of the beginning of the SURE intensives were reached, and then calculating trajectories and wet deposition statistics until fields corresponding to the dates and times of the ends of the SURE intensives were reached. The SURE intensive periods are identified in Sec. 3.3.1 of this report.

### 3.3 FIELD OBSERVATIONS OF PREDICTED VARIABLES

The data base used to compare against model predictions was a combination of atmospheric concentrations of  $\text{SO}_4^-$  and  $\text{SO}_2$  and wet  $\text{SO}_4^-$  deposition. Data collected over a three-year period (1978, 1980, and 1981), were considered in this study. These data were screened for completeness, and monthly and seasonal averages were computed for

statistical comparison with ASTRAP predictions. The data bases and screening procedures are described in the following sections.

### 3.3.1 Monthly Mean Air Concentration and Monthly Wet Deposition Data

The air-concentration measurements used to evaluate model performance came from SURE. Measurements were obtained from 54 ground-based air-quality stations: 9 Class I sites and 45 Class II sites. The Class II sites operated on intermittent schedules during seasonally representative sampling months (six intensive periods) beginning in August of 1977 and ending in October of 1978 (continuous 24-hour samples were collected during the intensive periods). The Class I sites operated on continuous schedules from August 1977 through June 1979 and had more stringent siting requirements than the Class II sites (continuous 3-hour samples were collected during the Class I sampling period). More complete descriptions of the SURE data base can be found in the literature (Mueller and Watson 1981; Hidy and Mueller 1981; EPRI 1979, 1982, and 1983). Four months of data collected in 1978 were available for use in our study. The sampling periods included January 10 through February 9 (31 days of winter samples), April 3 through May 2 (30 days of spring samples), July 1 through July 31 (31 days of summer samples), and October 1 through October 31 (31 days of fall samples). Monthly averages were computed for the 3-hour samples (Class I sites) and the 24-hour samples (Class II sites).

The locations of these sites are shown in Fig. 3.4. The site name, coordinates, and average monthly  $\text{SO}_4^{=}$  concentrations (January, April, July, and October) are given in App. D, Table D.1. These data, along with the  $\text{SO}_2$  data (see App. D, Table D.2), obtained from EPRI in February 1985, are in the most recent version of the SURE data base. They were used in preference to an earlier version of the data used in the MOI work, because some additional hourly data for computation of monthly averages were available and because some of the site data were missing in the MOI data base. A 75%-data-capture screening criterion was used to determine representative monthly averages. (The percent of data capture is computed by dividing the number of valid data points obtained by the total number of possible data points during the sampling period. At least 23 valid days were required to compute a valid monthly mean.) The data-capture values for each site are given in App. D, Table D.1. The cumulative frequency plots and histograms prepared to aid the data screening process are contained in App. D.

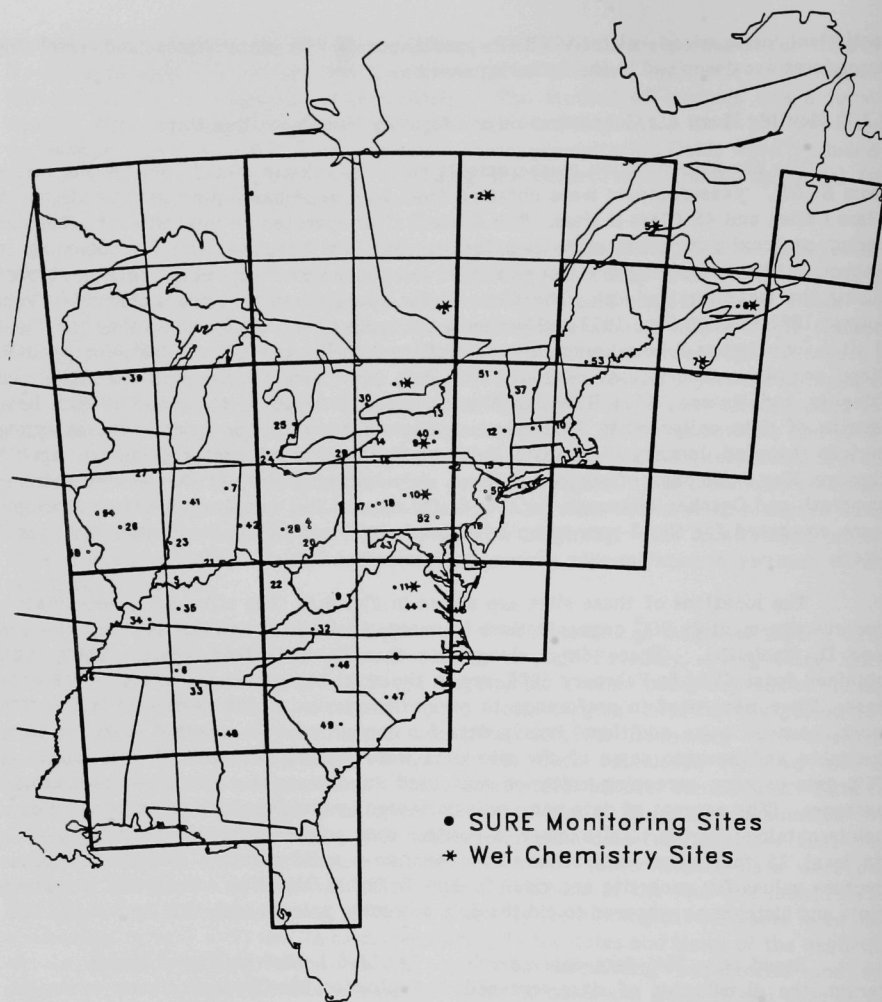
Based on a 75%-data-capture criterion, the distribution of sites screened for model evaluation is given in Table 3.1.

In addition to the measured air concentrations of  $\text{SO}_4^{=}$  and  $\text{SO}_2$  from the SURE network in 1978, a limited amount of data on wet sulfur deposition (as equivalent  $\text{SO}_4^{=}$ ) was also available for comparison with ASTRAP predictions. Most of the monitoring sites (eight) were from the Canadian CANSAP network (sampled monthly), but data were also

**TABLE 3.1 Distribution of Model Evaluation Air-Concentration Monitoring Sites by Month**

Month	Ambient $\text{SO}_2$	Ambient $\text{SO}_4^{=}$
January	41	30
April	38	30
July	37	46
October	40	41





**FIGURE 3.4 Sulfate and Sulfur Dioxide Air Concentration and Wet Chemistry Monitoring Sites in 1978**



available from three MAP3S event sites. All 11 sites operated for only one month, July 1978. The locations of these sites are also shown in Fig. 3.4.

### 3.3.2 Seasonal Wet-Deposition and Ionic Concentration Data

Two years of information on wet-deposition chemistry collected from seven networks operating over eastern North America were available at the time our study was initiated. These data came from the ADS, the official repository of North America precipitation chemistry measurements. This system provides a convenient computer data base for retrieval of statistical summaries of deposition monitoring data. An additional four years of data from 1982 through 1985 have just recently become available. Data were collected on an event, daily, weekly, or monthly sampling protocol. The deposition amounts were derived from observations in seven networks: (1) the National Atmospheric Deposition Program (NADP), (2) Utility Acid Precipitation Study Program (UAPSP), (3) MAP3S, (4) Air and Precipitation Monitoring Network (APN), (5) CANSAP, (6) Acid Precipitation in Ontario Study-Daily Network (APIOS-D), and (7) Acid Precipitation in Ontario Study-Cumulative Network (APIOS-C). Sampling protocols were event for MAP3S; daily for UAPSP, APIOS-D, and APN; weekly for NADP; monthly for CANSAP; and 28-day for APIOS-C.

The locations of the wet deposition sampling sites for the eight seasons of sampling, 1980 and 1982, are shown in Fig 3.5. Sites are numbered sequentially on a grid-by-grid basis. Those sites identified with an asterisk indicate co-located sampling. Table D.3, App. D, gives the site identifiers and names, coordinates, site completeness rating (discussed below), and wet  $\text{SO}_4^{2-}$  deposition fluxes. The same table for precipitation-weighted ionic concentrations (PWICs) is given in App. D, Table D.4, along with a sample calculation for computing PWIC.

To ensure a meaningful comparison between the model predictions and the observational field, a data screening procedure was instituted. The quality of the data was based on parameters measuring the completeness with which the sampling data were collected. Five different measures were used to determine data completeness. These parameters are defined in Table 3.2, along with the criteria used to rate or classify the samples. Three levels of completeness were assigned: Class A was given to sites with the highest rating, Class D to sites with the lowest rating. In addition to completeness, the degree to which a site was representative of its region was also considered as a screening criterion. Because of the limited amount of information on the effects of local sources on precipitation chemistry and the difficulty of obtaining sufficient data on sampling site characteristics that might influence the regional representativeness of the samples, data screening was based solely on data completeness.\*

---

\*Data screening based on a less stringent data-completeness criterion (Class B sites not screened) and on a regional representativeness criterion was used for the International Sulfur Deposition Model Evaluation (ISDME) study (Clark et al. 1987a). A comparison with our study shows that we screened fewer sites for each season in 1980. The distribution of sites in the ISDME model evaluation data base are as follows: winter 38, spring 46, summer 45, and autumn 42.

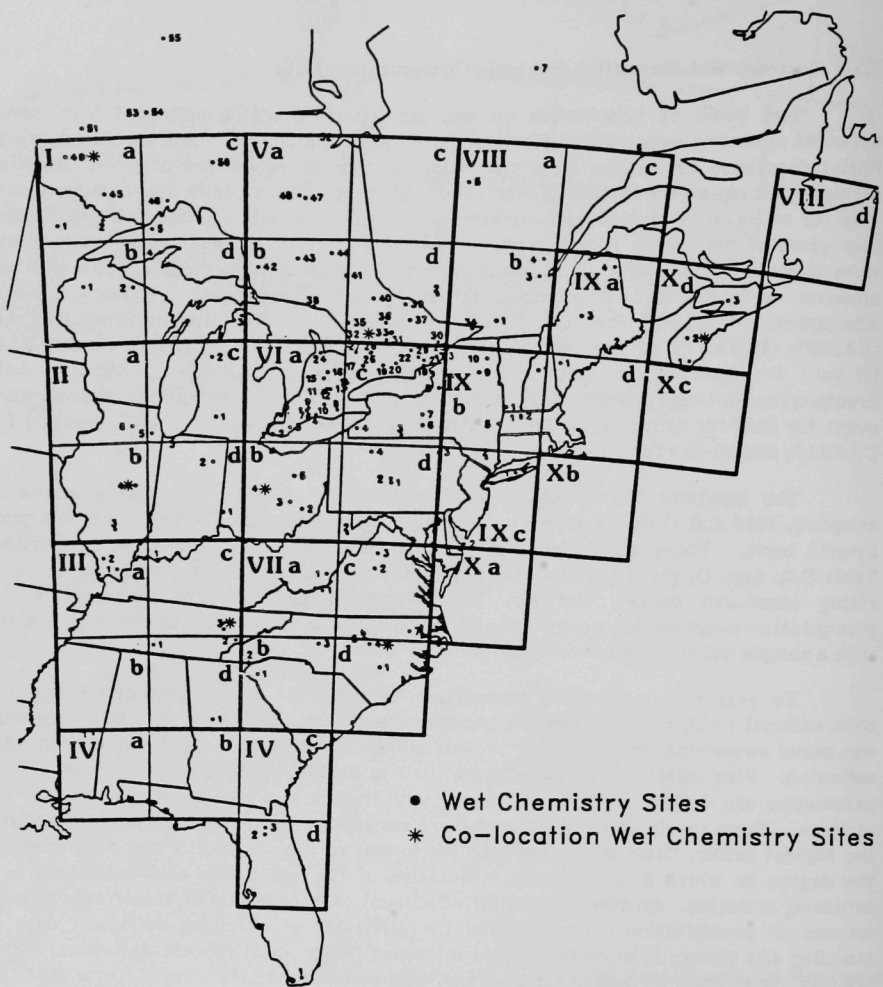


FIGURE 3.5 Wet Chemistry Monitoring Sites in 1980 and 1981

Based upon the completeness screening criteria given in Table 3.2, the distributions of observational sites for each season, for each year and for both years combined, are given in Table 3.3. The ASTRAP model performance evaluation was based on comparing model predictions to Class-A-rated monitoring sites. The number of sites passing the screening procedure ranged from 25 in winter 1980 to 70 in summer 1981. The percentage of data screened from the seasonal comparisons with ASTRAP ranged from 44% in summer 1981 to 71% in winter 1981.

### 3.4 BACKGROUND ADJUSTMENTS

Estimates of background concentrations of pollutant species in air masses that will later pass over man-made pollutant source areas are necessary in the regional-scale

**TABLE 3.2 Wet Sulfate Sampling Data Completeness Screening Criteria (%)**

Variable	Data Screening Class <sup>a</sup>		
	A	B	C
PCE	$\geq 80^b$ ( $\geq 50$ ) <sup>c</sup>	$\geq 60^b$ ( $\geq 30$ ) <sup>c</sup>	$\geq 50^b$ ( $\geq 30$ ) <sup>c</sup>
PTPVS	$\geq 90$	$\geq 80$	$\geq 60$
PCL	$\geq 95$	$\geq 90$	$\geq 90$
PVSL	$\geq 80$	$\geq 80$	$\geq 60$
PSMPV	$\geq 80$	$\geq 80$	$\geq 60$

Definitions of Screening Variables	Formula	ADS Variable Name
PCE = Percent collection efficiency	$\frac{\text{Total sample volume (converted to depth)}}{\text{Total precipitation depth}}$	SO <sub>4</sub> -PCE
PTPVS = Percent total precipitation associated with valid sample	$\frac{\text{Precip. depth assoc. with valid samples}}{\text{Total precipitation depth measurement during valid precipitation coverage}}$	SO <sub>4</sub> -PTP
PCL = Percent precipitation coverage length	Percent of the annual period for which valid precipitation coverage (in days) is available	SO <sub>4</sub> -PPCL
PVSL = Percent valid sample length	Percent of the annual period (in days) when valid sampling occurs	SO <sub>4</sub> -PVC
PSMPV = Percent of samples with measured precipitation that are valid	Fraction of the sampling periods during which precipitation is known to occur that resulted in valid precipitation chemistry data	SO <sub>4</sub> -VSMP

<sup>a</sup>Observations that do not meet any of the above criteria are assigned to Class D.

<sup>b</sup>NADP, MAP3S, and UAPSP networks (same percentage, as indicated, for all seasons).

<sup>c</sup>Winter PCE for CANSAP, APN, and APIOS networks.

**TABLE 3.3 Number of Sites that Meet Screening Criteria**

Season	Year	Data Screening Class				Total
		A	B	C	D	
Winter	1980	25	16	4	22	67
Spring	1980	29	16	5	25	75
Summer	1980	38	12	6	28	84
Autumn	1980	47	25	6	43	121
Total	1980	139	69	21	118	347
Winter	1981	37	18	24	37	116
Spring	1981	57	17	27	26	127
Summer	1981	70	20	16	18	124
Autumn	1981	66	23	26	15	130
Total	1981	230	78	93	96	497
Total	1980/ 1981	369	147	114	214	844

products. Not properly accounting for these background levels can result in systematic underprediction of observed concentration levels. Recognizing this, Workgroup 2 investigators during the MOI model innercomparison and evaluation work (MOI 1982) estimated an annual background level of 2 kg  $\text{SO}_4^-/\text{ha}$ , which was used to adjust predictions of three of the eight models evaluated in the study (some models considered emissions over a wider area, and thus the definition of "background" varied). However, no data documentation was provided in support of the level chosen. (The Canadian modelers felt that a background adjustment to their predictions was necessary and that the 2 kg  $\text{SO}_4^-/\text{ha}$  was a reasonable level to use.)

Nonanthropogenic (natural), uninventoried anthropogenic (man-made), and transported intercontinental (natural and man-made) sources of sulfur emissions contribute to continental background concentration and deposition flux levels that cannot be directly accounted for in regional scale model predictions. To minimize any systematic bias in comparing ASTRAP model predictions with observations, a procedure was developed to estimate a representative background level to be added to the model predictions. Three steps were followed: (1) identification and determination of the relative magnitude and spatial distribution of sources that might contribute to background levels, (2) examination of the spatial distribution of inventoried sources relative to natural sources (to determine if downwind sampling sites can be used as representative continental background sites), and (3) consideration of upwind sites outside the inventoried source region for establishing a representative background level.

Natural background emissions are produced by both biogenic and nonbiogenic sources. Biogenic sulfur emissions come from terrestrial (soils, crops, and natural vegetation) and oceanic (tidal and innertidal areas and nutrient-rich areas) regions where vegetative and microbial processes are active. Nonbiogenic sulfur emissions come primarily from geothermal sources such as volcanoes. On a nationwide basis, terrestrial sulfur emissions in the continental United States are an estimated 200,000 metric tons per year (t/yr) (NAPAP 1985). Oceanic emissions reaching the continental United States are nearly equivalent to terrestrial emissions at around 280,000 t/yr. Because of the relatively large spatial and temporal inhomogeneity of geothermal emissions, a representative estimate of the magnitude of these emissions is difficult to give. However, measurements taken during the Mt. St. Helen eruption in the spring of 1980 indicate that it contributed a smaller amount of sulfur to the atmosphere (e.g., ~80 t of  $\text{SO}_2$  with  $\text{SO}_4^-$  concentrations averaging 110 ppm over a 16-day period; Stöiber et al. 1980) than other volcanic eruptions (e.g., Irazu, Cost Rica, 1963; Pacaya, Guatemala, 1965) and biogenic sources.

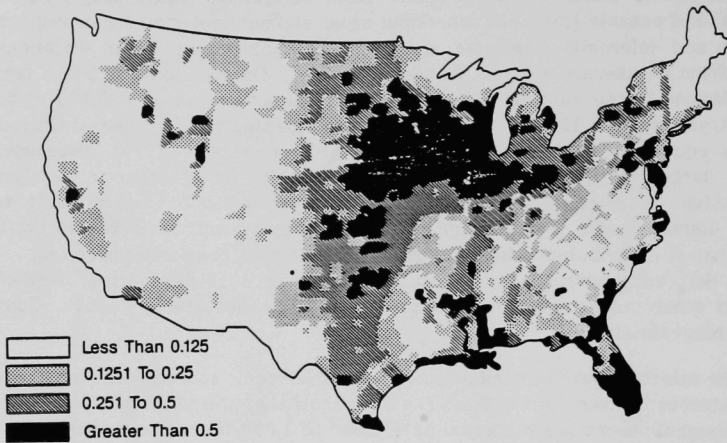
The relative spatial distribution of natural biogenic sources and man-made sulfur emission sources is shown in Fig 3.6. The data show that the range of emission densities for anthropogenic sources are approximately 300 to 1,000 times the corresponding range of emission densities for natural sources. Due to the spatial distribution of these sources and their relative magnitudes, representative natural sulfur background levels determined from samples located in North America would not be feasible, at least within the scope of this project. (It is too difficult to partition contributions from natural versus man-made sources without techniques such as a detailed elemental sample analysis based upon principal components.) Therefore, data collected from upwind monitoring sites in the global trends network (GTN) west of the source inventory region were examined to determine representative background level(s). The GTN provided a data base from sampling sites that were not near continental anthropogenic and nonanthropogenic sources.

The GTN has a total of twelve sites worldwide where precipitation chemistry samples are currently being collected. Figure 3.7 shows the location of these sites, in addition to three sites that have been closed (early 1981 and 1983, and late 1982).<sup>\*</sup> The sites identified as new were started after 1981, the period that followed the data base used for our comparisons. Average annual  $\text{SO}_4^-$  concentration levels measured at nine of the twelve currently operating GTN sites, along with the eastern North America average annual level, are given in Fig 3.8. A sufficient period of record for the three new stations (Kruger National Park, Cape Point, and Torres) was not available at the time of preparation of this report. The data show that the annual average concentration of  $\text{SO}_4^-$  over the high-density source region of eastern North America is over 10 times greater than that over remote regions of the world.

---

<sup>\*</sup>Background on the evolution of the U.S. precipitation chemistry monitoring program from the establishment of U.S.-world meteorological organization baseline sites (1972) to the establishment of GTN (1982) can be found in Dayan et al. (1985).

(a) Natural Sulfur Emissions (kg/ha/yr)



(b) Man-Made Sulfur Emissions (t/km<sup>2</sup>)

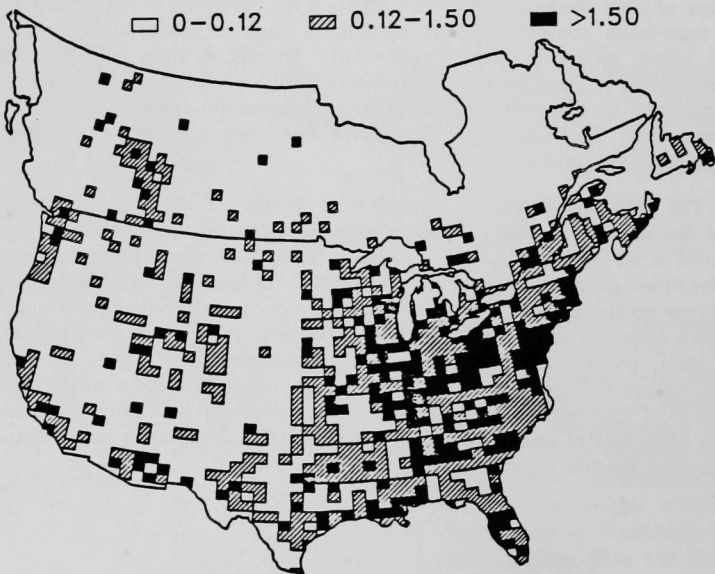
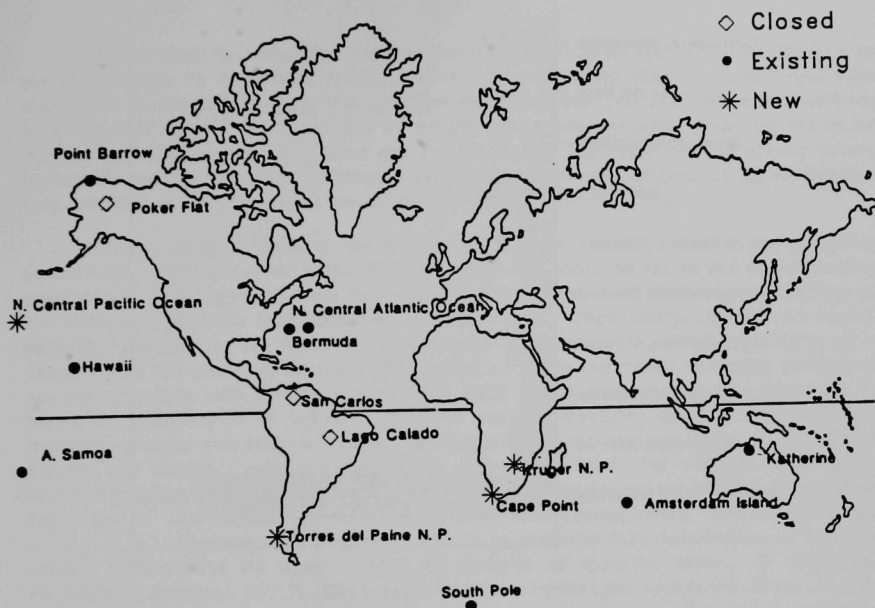


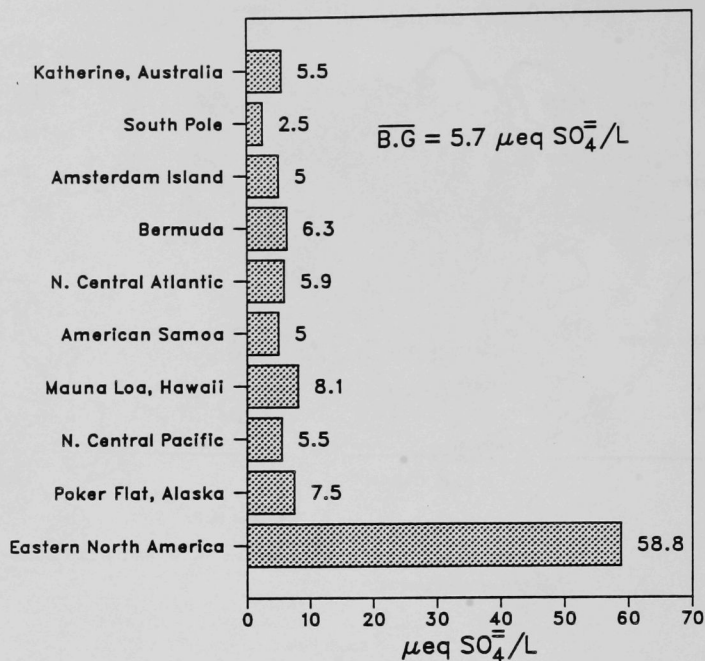
FIGURE 3.6 Natural (NAPAP 1985) and Man-Made (Wagner et al. 1986) Sulfur Emission Densities



**FIGURE 3.7 Global Trend Network Sites (Dayan et al. 1985)**

Five GTN sites, three in the Pacific and two in Alaska (see Fig. 3.7), were selected as being the most representative for approximating background levels in North America. Because of the relatively small amounts of precipitation at Point Barrow, Alaska, data from this site were not considered; and because of its inland location in central Alaska (70 km from Fairbanks), the data collected at Poker Flat were also not considered. That left measurements at three sites -- Mauna Loa, Hawaii; American Samoa; and the North Central Pacific -- for determination of a background level. Seasonal variations of  $\text{SO}_4^-$  ionic concentrations at each of these sites were not significant. The annual average of  $6 \mu\text{eq SO}_4^-/\text{liter}$  was determined to be a representative background level. This concentration was added to the seasonal wet  $\text{SO}_4^-$  predictions at each sampling site (subsequent to adjustment by seasonal precipitation amount at each site). The site-specific background values are given in Table D.3 in App. D.





**FIGURE 3.8 Mean Annual Ionic Concentrations of Sulfate at Global Trend Network Sites**

#### 4 OBJECTIVES AND METHODS FOR MODEL PERFORMANCE EVALUATION

Users need to know that models can be applied with reasonable confidence to predict changes in deposition and/or ambient concentration that are associated with changes in meteorological conditions and source emissions. This confidence is of critical importance if models are to be used for establishing source-receptor relationships under current emission control conditions and for evaluating the effectiveness of any future options for emission-control strategy. Error evaluation and uncertainty quantification help establish confidence in the use of models for policy formulation.

Error, as used here, is the difference between true values and corresponding predictions. *Apparent error* is the difference between observed values and corresponding predictions. Thus, *apparent error* includes the effect of errors in observations (defined as the difference between true values and measurements). *Uncertainty*, applied to a model, generally pertains to the range of expected error between the model predictions of a variable and the observed values of that variable. Since the model ordinarily predicts a number of values with varying errors, both error and uncertainty must be described in terms of distributions of values. We seek various measures of those distributions, including means, standard deviations, differences among various parts of the model domain, and ideally, a representation of the distribution itself. For example, *uncertainty* could be quantitatively expressed in terms of the joint conditional probability distribution for a set of true values  $x_1, x_2, \dots, x_n$ , given the corresponding model predictions  $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n$ . The task of model evaluation is to estimate that distribution, or at least certain measures of it, based in part on samples of *apparent error*. It should be emphasized, however, that *uncertainty* distributions reflect our knowledge of the model's performance and thus are subject to change as increased information becomes available. For example, we might expect that the mean square error between a given set of predictions and true values would become better defined (i.e., the variance about the expected value would become smaller) as more data become available for evaluation of the model.

As noted in the introduction to this report, a Bayesian (or Monte Carlo or some other) statistical theory is required to estimate the probability distribution described in the previous paragraph. Because of project and data constraints that we encountered (i.e., the treatment of the distribution of observational error) in the development of the Bayesian theory, our initial primary objective of developing an analytical framework for science-policy uncertainty quantification could not be carried through. However, we do believe that the objectives of the methodology for evaluating model performance described in the following paragraphs will provide some new and useful insights on understanding the characteristics, including magnitude and spatial patterns, of the *apparent error* in model predictions.

Our general objective is to develop a better understanding of the performance characteristics and *apparent error* of a long-term regional transport and deposition model. We intend to determine how well the predictions from the "standard" version of the ASTRAP model compare with corresponding observations. Another goal is to discern and quantify differences in spatial and temporal patterns in seasonal observations and

predictions. With respect to temporal patterns, we are interested in (1) how well the relative magnitudes of the maximum seasonal  $\text{SO}_4^-$  deposition and  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations (DAC) are reproduced in time, (2) differences in performance found in seasons of the same year (intraseasonal performance), and (3) differences in performance found in seasons of separate years (interannual performance). With respect to spatial patterns, we are interested in (1) the location and magnitude of maximums, and (2) the location, orientation, shape, and gradient in the DAC contours.

We also intend to find out, in an *operational* sense, how the relative performance of several parameter-adjusted versions of ASTRAP change over time and space. The issue of why, in a *diagnostic* sense, a model performs well or poorly is of key importance to model component and model data-base development and improvement. A *diagnostic* approach that employs, for example, sophisticated global sensitivity analysis is beyond the scope of this particular study. The results from this study should, therefore, be viewed from an application or *operational* perspective on evaluating and comparing models and model sensitivity rather than a *diagnostic* or research-oriented perspective on evaluating model performance and model component improvement.

This section of our report outlines the basis for selecting the performance evaluation measures and data analysis methodology. The selected statistical performance measures and methods are then described. With this foundation, the results from the model performance and sensitivity evaluation are presented and discussed (Sec. 5). Residual and scatter error patterns are presented with the aid of residual histograms, scatter plots, and time-series graphs along with univariate and difference statistical performance measures. The model performance sensitivity patterns resulting from internal model parameter variation are shown, and the relative and absolute magnitudes of model error are discussed. The decomposition of mean square error into its systematic and unsystematic components is used to quantify the potential magnitude of error reduction that can be achieved with the adjustment of model parameterization. The spatial, temporal, and bias components of error are quantified through decomposition of parametric statistical measures. Spatial patterns in model predictions and observations are displayed through interpolation among data points with least-squared regression and simple and universal kriging. Finally, some potential factors that may influence apparent model performance are investigated. They include the sampling protocol, geographic region, pairing observations and predictions as precipitation-weighted ionic concentrations versus mass deposition flux, and spatial scale of data aggregation.

#### 4.1 SELECTION OF PERFORMANCE EVALUATION MEASURES AND METHODS

Several factors were considered in selecting the statistical evaluation measures to be used in meeting our study objectives. The major factors included (1) the spatial and temporal resolution of the ASTRAP model and associated data bases, (2) the availability of observations of sufficient quality, quantity, and spatial distribution, and (3) the potential applications the model results would serve. Each of these factors constrains the applicability and interpretability of statistical performance measures. For example, the coarse spatial ( $\approx 130$  km) and temporal ( $\approx 1$  month) resolution of ASTRAP and

the sparse distribution of the precipitation chemistry data can hamper interpretability of standard parametric distributional statistics. The intended model applications can similarly influence the selection of performance measures to be used in the model evaluation. If it is important to replicate the upper end of the frequency distribution of observed short-term (1-hr to 24-hr averages) ambient concentrations, then distributional statistical measures, such as bias, correlation, and cumulative frequency plots, are important. On the other hand, if it is important to replicate spatial patterns in observed intermediate (monthly to quarterly periods) to long-term (annual period) deposition fluxes or ionic concentration fields, then graphical pattern recognition techniques (e.g., contour plots) and descriptive spatial statistics (e.g., decomposition of mean square error) are important.

Another issue in the choice of performance measures is determining the interpretability or significance of the computed statistic. In "significance tests" (e.g., "p value" confidence intervals, Mason-Whitney test, Bartlett's test), the data are assumed to be statistically independent, but this is not the case in our study because of the spatial correlation inherent in our data set. Also, significance testing is not recommended when the extent of violation of the assumptions underlying the particular test is unknown, and the power of the test is as much a function of the number of data points, the sample distribution, and the test itself, as it is a function of the true relationships contained in the data (Wilmott 1981). Therefore, these significance tests are not used to determine the probability that the differences between predictions and observations were not obtained by chance. In addition, parametric measures such as mean bias error, variance, and correlation cannot alone reveal the true spatial characteristics of the data.

As a result of the above considerations, it was determined that standard model evaluation methods, such as those recommended by the American Meteorological Society (AMS) on quantifying and communicating model uncertainty (Fox 1984), cannot be used exclusively. An approach was selected that combines the more traditional distributional measures from the AMS with some new measures that allow for decomposition of error components and a more robust analysis of spatial, temporal, and bias error. Some of these measures are described briefly in the next section and more completely, with full mathematical detail, in App. E.

## **4.2 DESCRIPTION OF PERFORMANCE EVALUATION MEASURES AND DATA ANALYSIS METHODOLOGY**

Statistical measures for evaluating model performance can basically be categorized as parametric or nonparametric techniques. Parametric statistics assume the data can be fit to some standard distribution function, such as a Gaussian or normal distribution. Nonparametric data analysis techniques (such as the bootstrap method from Efron and Gong 1983) and numerical methods (such as principal component, numerical correlogram, time-series, sensitivity, spectral, and empirical orthogonal function analysis) were considered but were beyond the scope of this particular study. Therefore, our study was restricted to the use of parametric statistics. The parametric statistical measures considered can be grouped as either graphical analysis or descriptive analysis techniques. The graphical techniques listed in Table 4.1 can represent in a meaningful way a model's ability to accurately characterize observational data. This representation,

in turn, provides a valuable means to identify spatial and temporal patterns of model performance and the sensitivity of these patterns to perturbations in the model's empirical or theoretically derived parameters. The descriptive measures listed in Table 4.2 express the patterns of performance identified through graphical analysis in quantitative terms. The measures help a model evaluator to discern and compare levels of acceptable performance among several models or model versions, especially if sensitivity to adjustments in model parameters is a component of the model evaluation study.

Descriptive measures are of two types: univariate measures and difference measures. Univariate (single variable) measures can be used alone or in combination with graphical displays or difference measures\* to express means, mean squares, and variances in observations and model predictions. Difference measures (two variables) are of three kinds: arithmetic indices, nondimensional indices, and logarithmic indices. Arithmetic indices can be used to express bias, variance, correlation, and mean or root mean square error between observations and model predictions. Nondimensional indices are valuable measures when the comparison of the performance of different models and/or at different time periods is important. These measures can also be combined with arithmetic indices or graphical displays to quantitatively characterize the patterns and components of error. Finally, logarithmic indices are useful, in combination with nondimensional indices, as an aid in ranking model performance. These indices are also an important component of the Bayesian uncertainty theory as developed under this project.

The following discussion describes some of the graphical techniques and the principal descriptive statistical methods used in evaluating ASTRAP model performance.

#### 4.2.1 Pattern Recognition through Data Display Techniques

The most commonly used data displays in model evaluation studies are scatter plots of model-predicted (P) versus "reliable" field-observed (O) variables and frequency histograms. The relationship between P and O can be well represented by scatter plots in combination with descriptive performance measures. Scatter plots are particularly helpful in uncovering underlying systematic differences between P and O as well as troublesome extremes. The spatial dependence of model performance, if such

**TABLE 4.1 Graphical Statistics Pattern Recognition Techniques**

---

Scatter plots
Normalized bias-scatter error plots
Fractional bias-scatter error plots
Residual histograms
Cumulative frequency distributions
Box plots
Observation/prediction histograms
Residual vs. prediction histograms
Time-series plots
Contour plots
• Kriged
• Unkriged

---

\*When used with difference measures, univariate measures can be used to decompose error into its spatial, temporal, and bias components.

**TABLE 4.2 Descriptive Statistical Model Performance Measures and Indices***Univariate Measures*

$\bar{O}$ , $\bar{P}$	- Mean Observation and Mean Prediction
$\sigma_o$ , $\sigma_p$	- Standard Deviation of Observations and Predictions
$CV_o$	- Coefficient of Variation of Observations
$CV_p$	- Coefficient of Variation of Predictions
$MSO$	- Mean Square Observation
$MSP$	- Mean Square Prediction
$MSTO$	- Mean Square Temporal Observation
$MSTP$	- Mean Square Temporal Prediction
$MSSO$	- Mean Square Spatial Observation
$MSSP$	- Mean Square Spatial Prediction

*Difference Measures**Arithmetic Indices*

## 1. Bias Error

MBE	- Mean Bias Error
MABE	- Mean Absolute Bias Error
MRPP	- Mean Residual-Prediction Product
FABE	- Fractional Average Bias Error
MRR	- Mean Residual Ratio
RMR	- Residual Mean Ratio

## 2. Correlation and Variance

$R_1$	- Coefficient of Correlation between Observations and Predictions
$COV_1$	- Covariance
$R_2$	- Coefficient of Correlation between Residuals and Predictions
$COV_2$	- Residual Covariance
FSE	- Fractional Scatter Error
VAR	- Variance
STD	- Standard Deviation
$CV_r$	- Coefficient of Variation of Residuals
RPRS	- Relative Prediction to Residual Scatter



TABLE 4.2 (Cont'd)

---

 3. Mean and Root Mean Square Error and Error Decomposition

MSE	- Mean Square Error
RMSE	- Root Mean Square Error
MSE <sub>u</sub>	- Mean Square Error Unsystematic
MSE <sub>s</sub>	- Mean Square Error Systematic
	MSE <sub>a</sub> - Additive Systematic MSE
	MSE <sub>p</sub> - Proportional Systematic MSE
	MSE <sub>i</sub> - Interdependent Systematic MSE
MSTE	- Mean Square Temporal Error
MSSE	- Mean Square Spatial Error
RDMSE	- Relative DMSE (as percent of the normalized MSD)

## Logarithmic Indices

MSLE	- Mean Square Logarithmic Error
	MSLTE - Mean Square Logarithmic Temporal Error
	MSLSE - Mean Square Logarithmic Spatial Error
MLE	- Mean Log Error
VLE	- Variance Log Error
SLE	- Standard Deviation Log Error
GMSE	- Geometric Mean Square Logarithmic Error
GMLE	- Geometric Mean Log Error
GSDL	- Geometric Standard Deviation Log Error

## Nondimensional Indices

IOA	- Index of Agreement
DMSE	- Dimensionless Mean Square Error
FABE	- Fractional Average Bias Error
FSE	- Fractional Scatter Error
NMBE	- Normalized Mean Bias Error
NSE	- Normalized Scatter Error
RPRS	- Relative Prediction to Residual Scatter
RDMSE	- Relative DMSE (as percent of the normalized MSO)
RSI	- Rank Score Index

---

dependence exists, can be displayed on scatter diagrams by identifying groups of data points by region or area. Frequency histograms are useful in displaying the distribution and degree of bias in model predictions (residual histograms) and are useful in showing how that bias varies over different time periods. The relative distributions of predictions versus observations and the shape of those distributions can also be displayed with frequency histograms.

The analysis of time-series plots can provide a clear picture of temporal patterns in P and O. For example, a time-series plot can show the relative model performance during climatologically different time periods and reveal the bias and temporal error in



that performance. However, these plots can be difficult to interpret when P and O trends are highly variable. The intraseasonal and interannual differences in observation and predictions and the size of the residual can be visually displayed and easily identified with these plots. Observations from an individual monitor and averaged observations from multiple monitors spatially representing the model resolution-size receptor grid can be shown and evaluated. The time-history of precipitation amounts can also be displayed and correlated with the residual time-history. Until recently, only a limited historical record of reliable observational data (e.g., wet  $\text{SO}_4^-$  deposition) was available to model evaluators (only two years were available at the beginning of this study). As a result, time-series plots have not been used often for evaluating long-term predictions from regional-scale transport and deposition models. The reliable data available since 1980 have made the use of time-series evaluation of seasonal or monthly deposition predictions more feasible. Previous studies using time-series plots have been done with short-term predictions (averaging  $\leq 24$  hr) for short time periods (a month or so duration). McNaughton et al. (1980, 1981) used time-series plots in evaluating the performance of the Regional Air Pollutant Transport (RAPT) model's predictions against SURE daily  $\text{SO}_4^-$  air concentration observations.

The relative bias and scatter in model predictions can be visually displayed with fractional and normalized error plots. Fractional error (FE) plots provide visual information about the overall goodness of fit between observations and predictions, in addition to revealing patterns in bias and scatter error. Fractional scatter error (FSE) in model predictions are plotted on the ordinate, while fractional average bias error (FABE) in those predictions are plotted on the abscissa. The fractional bias and scatter are computed as follows:

$$\text{FABE} = \frac{2 (\bar{O} - \bar{P})}{\bar{O} + \bar{P}} \quad (4.1)$$

$$\text{FSE} = \frac{2 (\sigma_o - \sigma_p)}{\sigma_o + \sigma_p} \quad (4.2)$$

where:

$\bar{O}$  and  $\bar{P}$  = the overall average observation and prediction, and

$\sigma_o$  and  $\sigma_p$  = the standard deviation of observations and predictions.

FABE measures how well, on the average, a model estimates observational fields. If FABE is less than +0.67 and more than -0.67, model predictions are within a factor of two of observations. As FABE approaches  $\pm 2.0$ , the model is producing extreme over- or underpredictions. FSE measures how well, on the average, a model estimates the scatter among observations. It represents the difference between the standard deviations of observations and predictions. If FSE is less than +0.67 and greater than -0.67, the scatter in model predictions is within a factor of two of the scatter in observations. As FSE approaches  $\pm 2.0$ , the model predictions of scatter in the observations are extreme. Fractional error plots have been used primarily to describe the performance of short-term Gaussian air quality dispersion models over local transport scales (Cox et al.

1985a, 1985b; Irwin and Smith 1984). The use of these plots, which employ fractional differences in means and standard deviations of observations and predictions, may not be appropriate for quantifying the bias and scatter error for a long-term regional-scale transport and deposition model. The reason is because of the spatial and temporal differences between short-term, local-scale air-quality models and long-term, regional-scale models (i.e., ASTRAP). The extreme end of the cumulative frequency distribution is important in evaluating short-term air-quality models, while cumulative deposition totals and mean concentrations are more important in evaluating long-term regional models. Therefore, FE plots are used in our evaluation of ASTRAP primarily to display sensitivity patterns in internal model parameter bias and scatter error.

The magnitude of the bias and scatter error, in addition to the patterns of sensitivity in this error, can be displayed with normalized error (NE) plots. The normalized scatter error (NSE) in model predictions is plotted on the ordinate while the normalized mean bias error (NMBE) is plotted on the abscissa. The normalized bias and normalized scatter error are computed as follows:

$$NMBE = \sqrt{\frac{\bar{r}}{(\sigma_b \cdot \sigma_p)}} \quad (4.3)$$

$$NSE = \sqrt{\frac{\sigma_r}{(\sigma_o \cdot \sigma_p)}} \quad (4.4)$$

where:

$$\bar{r} = \frac{1}{N} \sum_{i=1}^N O_i - P_i,$$

$$\sigma_r = \frac{1}{N-1} \sum_{i=1}^N (r_i - \bar{r})^2{}^{1/2},$$

$r_i$  = residual or difference between observation and prediction, and

$N$  = number of observation-prediction pairs.

The distance of each data point from the origin on an NE plot is proportional to the mean square error (MSE) in model predictions. Model mean biases and random variances are accounted for in the MSE. This measure is computed as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^N (O_i - P_i)^2 = \frac{1}{N} \sum_{i=1}^N r_i^2 \quad (4.5)$$

The mean bias portion and random variance portion of the MSE error is shown in the expression below, which can be shown to be equivalent to Eq. 4.5.

$$\text{MSE} = \frac{N-1}{N} \sigma_r^2 + (\bar{r})^2 \quad (4.6)$$

The spatial patterns in model performance and the *apparent error* in those patterns are determined through a geostatistical spatial extrapolation technique called kriging. A description of the kriging approach is given in Sec. 5.3.2 and App. L.

#### 4.2.2 Descriptive Statistical Performance Measures and Indices

As mentioned earlier, the majority of the statistical performance measures listed in Table 4.2 are described in detail in App. E. Some of the individual or combined measures or indices that play a key role in characterizing model performance are described here. These measures include the index of agreement (IOA), dimensionless mean square error (DMSE), relative DMSE, mean logarithmic error (MLE), variance logarithmic error (VLE), rank score index (RSI), systematic mean square error (MSES), and spatial/temporal/bias error components.

The IOA is greater than or equal to zero and smaller than or equal to one. It is defined by Willmott (1981) as follows:

$$\text{IOA} = 1 - \frac{\sum_{i=1}^N (P_i - O_i)^2}{\sum_{i=1}^N [ |P'_i| + |O'_i| ]^2} \quad (4.7)$$

where:

$P_i$  = the model prediction over grid cell  $i$ ,

$O_i$  = the model observation over grid cell  $i$ ,

$P'_i = P_i - \bar{O}$ ,

$O'_i = O_i - \bar{O}$ , and

$\bar{O}$  = the mean observation over all grid cells.

The right-hand side of Eq. 4.7 is the ratio of the mean square error and potential error. This index specifies the degree to which the predicted and observed deviations about the mean observation correspond. In a formal sense, it is not a correlation or association measure but rather a measure of the degree to which a model's predictions are "error-free" (assuming that the mean observation is "error-free"). The IOA is a standardized measure that provides a means to compare the performance of different models or model

versions, or to compare model performance under different atmospheric conditions or during different time periods. It accounts for both bias and scatter in model predictions but is more sensitive to scatter than bias in its indication of the relative amount of the apparent model error. The closer the index is to 1.0, the better the model performance.

The DMSE is defined by Hanna and Heinhold (1985) as follows:

$$DMSE = \frac{\sum_{i=1}^N (P_i - O_i)^2}{\bar{O} \cdot \bar{P}} \quad (4.8)$$

This measure, like IOA, accounts for both mean biases and the scatter or random variations in model predictions. Unlike IOA, DMSE seems to be more sensitive to changes in bias error than to changes in scatter error or variance (see discussion and data plots in Secs. 5.2.1 and 5.2.2). The smaller the DMSE, the better the model performance. As defined, the measure places more weight on higher concentrations because the prediction-observation differences are more likely to be the largest at the highest concentrations. Confidence intervals on the difference in DMSE between two models or model versions ( $DMSE_1$  and  $DMSE_2$ ) can be assessed with a Chi-square evaluation, if the expected differences ( $DSME_1 - DSME_2$ ) are normally distributed and  $DSME_1$  and  $DSME_2$  are independent. Because both of these conditions are not likely to be met, if the data set is small ( $N < 100$ ), a procedure for better defining confidence limits is desired. The bootstrap method developed by Efron and Gong (1983) provides a means to establish confidence intervals for small data sets. The computer-intensive requirements of this method alone, with limits imposed by project budget, prohibited the use of bootstrapping procedures in our study.

The relative DMSE is the total error expressed as a percent of the dimensionless mean square observation. It provides a less biased measure of error across short time periods (seasonal or monthly) than the relative MSE used by Fay et al. (1985) to quantify the "model" error for annual period(s) of deposition.\* The RDMSE is computed as follows:

$$RDMSE = \frac{MSE/(\bar{O} \cdot \bar{P})}{MSO/\bar{O}^2} \cdot 100\% = \frac{DMSE \cdot \bar{O}^2}{MSO} \cdot 100\% \quad (4.9)$$

where the numerator is the DMSE and the denominator is the dimensionless mean square observation. The dimensionless mean square observation is computed as the sum of the squared observations divided by the square of the mean observation.

The MLE and the VLE were measures derived primarily for use in the Bayesian model development (Ball 1986). Some of the parameters of more normally distributed

---

\*Variability of mean deposition across annual time periods is less than the variability across seasons (e.g., winter/summer).

data that resulted from taking logarithms made the use of these measures as bias and variance indicators attractive. These measures are computed as follows:

$$MLE = \frac{1}{N} \sum_{i=1}^N \ln \left( \frac{O}{P} \right)_i \quad (4.10)$$

$$VLE = \frac{1}{N-1} \sum_{i=1}^N \left[ \ln \left( \frac{O}{P} \right)_i - \frac{1}{N} \sum_{i=1}^N \ln \left( \frac{O}{P} \right)_i \right]^2 \quad (4.11)$$

A dimensionless index that combines bias- and scatter-measuring properties of the IOA, VLE, DMSE, and MLE would be a very useful measure for comparing the performance of different models or the performance of varying internal parameters of the same model. It would also provides a more robust measure for ranking model performance. This dimensionless RSI can be derived from Eqs. 4.7, 4.8, 4.10, and 4.11 as:

$$RSI = \frac{IOA (DMSE + MLE + VLE) + 1}{IOA} \quad (4.12)$$

An errorless model is indicated by an index of 1.0, while an RSI greater than 2.0 indicates poor model performance. Average model performance can be assumed with values somewhere between 1.65 and 2.00.

Willmott (1981 and 1982) suggests a means to decompose MSE into its systematic and unsystematic components. This provides a means of calculating the potential error reduction that can be achieved while studying the sensitivity of a model to variations in model-dependent variables. The systematic MSE measures reducible model and data-base uncertainty and is computed as:

$$MSE_s = \frac{1}{N} \sum_{i=1}^N (\hat{P}_i - O_i)^2 \quad (4.13)$$

where  $\hat{P}_i = a + b \cdot O_i$ , and  $a$  and  $b$  are the intercept and slope of the regression of  $P$  on  $O$ . An ordinary least-squares fit can be assumed under the proposition that  $P$  is linearly dependent on  $O$ . The assumption of  $P$  as the dependent variable and  $O$  as the independent variable is particularly important, since it implies that  $O$  is error-free and that all the error variance is contained within  $P$ . With very good observational data, this assumption is quite reasonable, although  $O$  is rarely, if ever, error-free (Willmott 1981).\*

---

\*Willmott also suggests that systematic error can be further decomposed into an additive, proportional, and interdependent component (see App. E), but the utility of such decomposition is not immediately known.

The unsystematic component of MSE can be assumed to be a measure of the potential accuracy (explained in the following text) of the model and data base and can be computed as:

$$MSE_u = \frac{1}{N} \sum_{i=1}^N (P_i - \hat{P}_i)^2 \quad (4.14)$$

or, more efficiently, as:

$$MSE_u = MSE - MSE_s \quad (4.15)$$

The error expressed in Eqs. 4.13, 4.14, and 4.15 can be more easily interpreted, in units of P and O, by taking the square roots of the MSE. With a good model, the systematic differences should approach zero, while the unsystematic differences approach the root mean square error:  $RMSE = [(RMSE_s)^2 + (RMSE_u)^2]^{1/2}$ . If the O and P differences described by  $RMSE_s$  can be described by a linear function, these differences should be relatively easy to dampen with simple model adjustments, for example, revisions to the model parameterization. In other words, without change or significant changes to the model's structure (governing equations), it should be possible to reduce the systematic portion ( $RMSE_s$ ) of the apparent model error. This implies that the unsystematic portion ( $RMSE_u$ ) can be interpreted as a measure of potential model accuracy (Willmott 1982).

Finally, MSE can be decomposed through analysis of variance into its bias, spatial, and temporal components. By taking the sum of squares within groups (over all similar time periods, e.g., winters, summers, etc.) at each receptor grid region, the mean square temporal error can be computed as follows:

$$MSTE = \frac{\sum_{i=1}^M \sum_{k=1}^{K_i} [O_{ik} - P_{ik} - \langle r_i \rangle]^2}{N - M} \quad (4.16)$$

By taking the sum of squares between groups over all receptor grid regions, the mean square spatial error can be computed as follows:

$$MSSE = \sum_{i=1}^M \frac{K_i [\langle r_i \rangle - \langle \langle r \rangle \rangle]^2}{M - 1} \quad (4.17)$$

where:

$O_{ik}$  = observation at receptor i during time period k,

$P_{ik}$  = prediction at receptor i during time period k,

M = cumulative total number of nonzero observation sites producing at least one observation-prediction pair,

$K_i$  = number of observations (seasons or months) at site  $i$ ,

$N = \sum_{i=1}^M K_i$  = total number of nonzero observation-prediction pairs,

$i$  = receptor index,

$k$  = time period index,

$$\langle r_i \rangle = \frac{1}{K_i} \left[ \sum_{k=1}^{K_i} O_{ik} - \sum_{k=1}^{K_i} P_{ik} \right]$$

= the mean residual at receptor averaged over all time periods, and

$$\langle\langle r \rangle\rangle = \frac{1}{N} \left[ \sum_{i=1}^M \sum_{k=1}^{K_i} O_{ik} - \sum_{i=1}^M \sum_{k=1}^{K_i} P_{ik} \right] \quad (4.18)$$

= the mean residual over the entire field of values.

Equations 4.16, 4.17, and 4.18 can now be used to decompose the MSE (Eq. 4.5) into its temporal, spatial, and bias components, as:

$$MSE = \frac{N-M}{N} MSTE + \frac{M-1}{N} MSSE + \langle\langle r \rangle\rangle^2 \quad (4.19)$$

The derivation, modification, and use of Eq. 4.19 is discussed in Sec. 5.3.1, Eq. 5.8, with results of the analysis of the ASTRAP prediction of wet sulfate deposition presented. Also discussed, with results, is the derivation of explained variance in terms of a spatial and temporal component. This measure of model performance is very useful, since the bias-induced observational error is corrected for in the expression derived for the explained variance in model predictions.



## 5 MODEL PERFORMANCE AND SENSITIVITY EVALUATION RESULTS

Regional-scale Lagrangian models such as ASTRAP have a relatively coarse spatial and temporal resolution. ASTRAP has a temporal resolution of one month and a spatial resolution of 100 to 130 km. The choices available for pairing of observations and predictions are restricted by the spatial and temporal resolution constraints of the model, the constraints of the observational data base (e.g., number of valid and representative sites and number of samples per site), and the error that can be introduced by data aggregation (time and space scales). Our comparisons were restricted by the number of data points available to unpaired data or data paired only in time. In addition to individual observation and prediction pairings in time (by season), several levels of time and space aggregation were examined. Spatial aggregation was taken on a unit-grid increment (30-40 km), 9-grid increment (300-390 km), and 36-grid increment (1,200-1,560 km) basis.\* Simple arithmetic averages were computed for each aggregation level. Observations and predictions were paired on a unit-grid basis, and performance results were reported on this basis. The effects that aggregating predictions and observations over large spatial scales has on model performance are discussed under special topics in Sec. 5.4. Temporal aggregation of seasons was taken over one year (four seasons), two years (eight seasons), and like seasons (four groups of two seasons), with a season being as previously defined in Sec. 3.

Our results are reported under four subject areas. The residual and scatter patterns in model performance are brought out with the use of scatter, time-series, and residual histogram plots (Sec. 5.1). A variety of statistical measures of bias and variance are used in describing the patterns observed with the data graphics. The evaluation of model sensitivity to variation in internal model parameters is covered next (Sec. 5.2). Fractional error and normalized error plots are used to display sensitivity patterns to variations in internal model parameters. We then provide an analysis of spatial patterns through the use of contour plots and decompose error into its bias, spatial, and temporal components (Sec. 5.3). Finally, in Sec. 5.4, a number of special topics are covered on factors influencing apparent model performance and interpretation of model results. The results should be viewed collectively because no single group of performance measures (e.g., descriptive difference statistics, nondimensional indices, graphical statistics) can describe all the aspects that are significant when judging how well a model characterizes observational fields. Even when viewed collectively, if the error associated with the computation of field sampling and data analysis is not well defined, the interpretation of model performance results, particularly the identification of why the model performs well or poorly, becomes exceedingly difficult.

---

\*Winds for ASTRAP are analyzed over an NMC grid cell, while precipitation is analyzed over a one-third NMC grid cell. The NMC grid cells (each composed of nine unit-grid increments) are displayed in the model evaluation grid, Fig. 3.1. The model evaluation region (MER) used in this study is composed of 360 of these unit-grids. The relative scale of the 9-grid (40 per MER) and the 36-grid (10 per MER) increments can also be seen in Fig. 3.1 (designated by small letters and roman numerals).

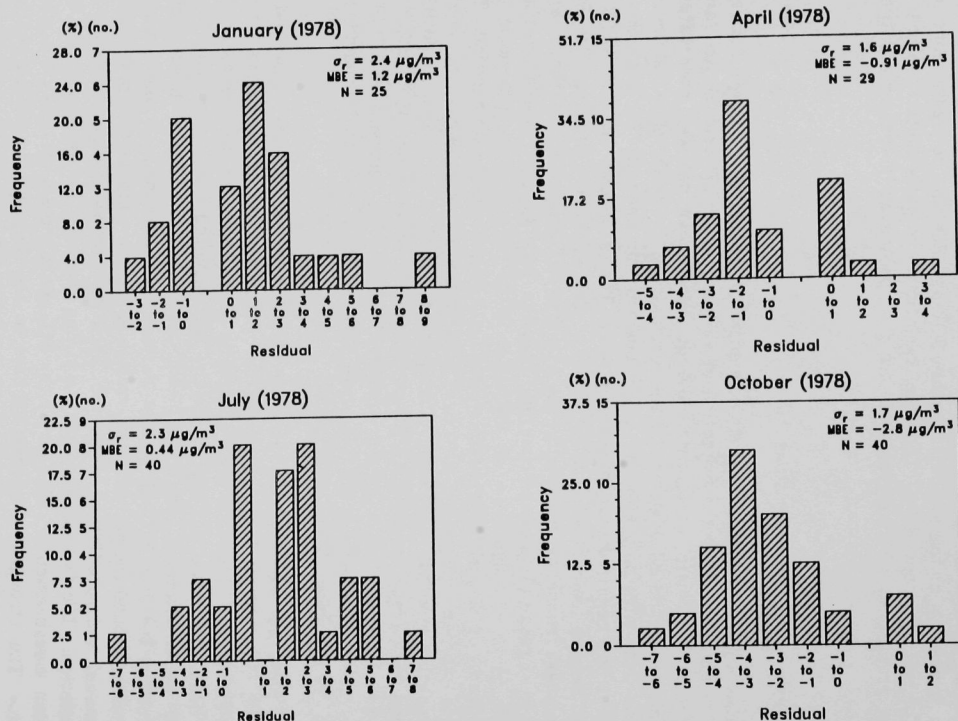
## 5.1 RESIDUAL AND SCATTER ERROR PATTERNS

Residual and scatter performance measures and graphical data displays can help to reveal some of the temporal patterns in model performance. These patterns are presented and discussed for the 1978 air-concentration data (Sec. 5.1.1) and the 1980 and 1981 wet-deposition data (Sec. 5.1.2). Temporal variations in residual patterns are illustrated with frequency histograms of monthly  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations and seasonal fluxes of wet  $\text{SO}_4^-$  deposition. Temporal variations and some spatial features of observations and predictions of these variables are then presented in scatter and time-series plots.

### 5.1.1 Monthly Average Air Concentrations

The frequency histograms of  $\text{SO}_4^-$  air concentration residuals are shown for each of four months in Fig. 5.1. With the exception of October, the  $\text{SO}_4^-$  residuals are fairly close to being normally distributed, with 60% of differences between observations predictions within one standard deviation of the residuals ( $\sigma_r$ ). The October distribution is skewed to the left, with only 30% of the residuals within one  $\sigma_r$ . The residuals in October show a strong negative bias (overprediction). The mean bias error in ASTRAP  $\text{SO}_4^-$  predictions ranged from  $\pm 0.4 \mu\text{g}/\text{m}^3$  in July to  $-2.8 \mu\text{g}/\text{m}^3$  in October. The  $\text{SO}_4^-$  residuals with highest frequency occurred in the 1 to  $2 \mu\text{g}/\text{m}^3$  range (24% of the time) in January, the -2 to -1 range (38% of the time) in April, the -1 to 2 range (40% of the time) in July, and the -4 to -3 range (30% of the time) in October. Figure 5.2 shows the frequency histograms of  $\text{SO}_2$  air concentration residuals over the same time period. The  $\text{SO}_2$  distributions approach normality, with over 60% of the  $\text{SO}_2$  residuals within one  $\sigma_r$  for all four simulation months. The January and April simulations have a slight positive bias. The mean bias error in ASTRAP  $\text{SO}_2$  predictions ranged from  $\pm 2.0 \mu\text{g}/\text{m}^3$  in October to 12.6 in January. The highest-frequency  $\text{SO}_2$  residuals occurred in the 10 to  $20 \mu\text{g}/\text{m}^3$  range (22%) in January, the 0 to 5 range (24%) in April, the 0 to 10 range (43%) in July, and the -10 to 0 range (48%) in October.

Scatter plots of monthly  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations are presented in Figs. 5.3 and 5.4. Perfect fit is indicated by the center dashed line, while predictions that are a factor of two under or over observations are indicated by the outer two dashed-dotted lines. The solid line represents the least-squared linear regression fit of the data. The slope, intercept, and correlation coefficient are shown on each plot. ASTRAP predictions for July  $\text{SO}_4^-$  air concentrations show the smallest and the most symmetrical scatter around the perfect prediction line. The January scatter plot shows a tendency for model underprediction, while the April plot shows a slight tendency for model overprediction. The October data show a strong tendency for overpredicting  $\text{SO}_4^-$  air concentration measurements. All four of the  $\text{SO}_2$  plots show a tendency for model underprediction. The strongest tendency appears in the January simulation, while a slight tendency to underpredict appears in the July simulation. The  $\text{SO}_4^-$  bias tendencies are further supported by a positive MBE of 1.2 for January and 0.4 for July, and a negative MBE of 0.9 for April and 2.8 for October. Likewise, the positive MBE for  $\text{SO}_2$  ranging from 2.0 (October) to 12.6 (January) supports the degree of positive bias observed



**FIGURE 5.1** Frequency Histograms of Monthly Sulfate Air Concentration Residuals ( $\mu\text{g SO}_4^-/\text{m}^3$ )

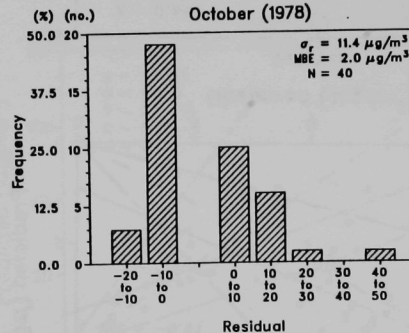
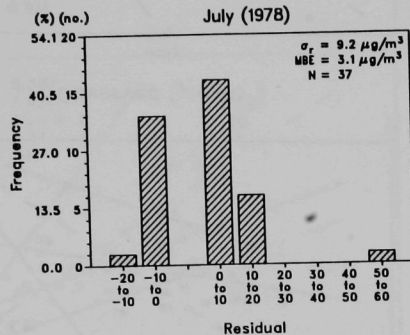
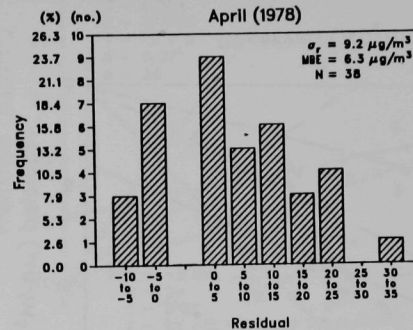
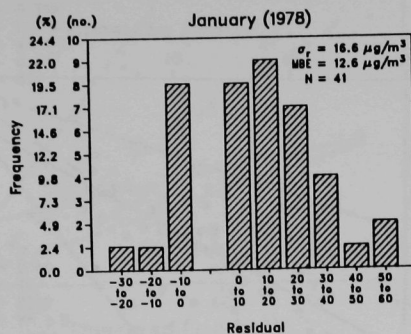


FIGURE 5.2 Frequency Histograms of Monthly Sulfur Dioxide Air Concentration Residuals ( $\mu\text{g SO}_2/\text{m}^3$ )

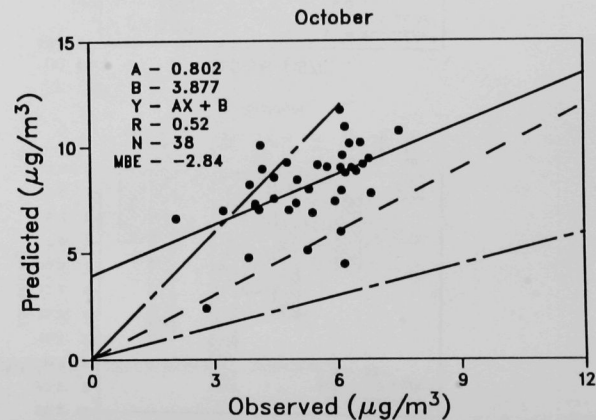
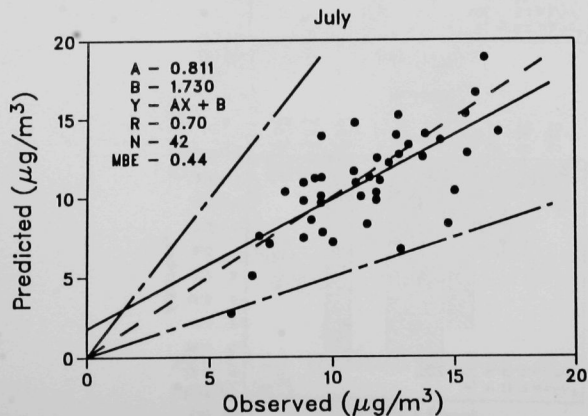
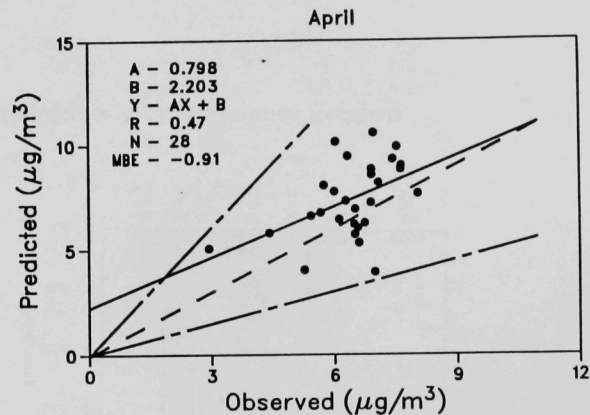
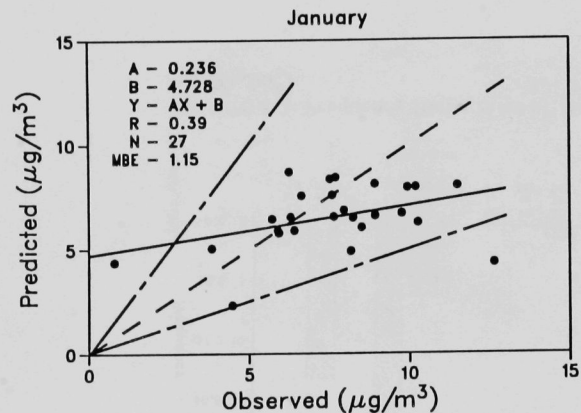
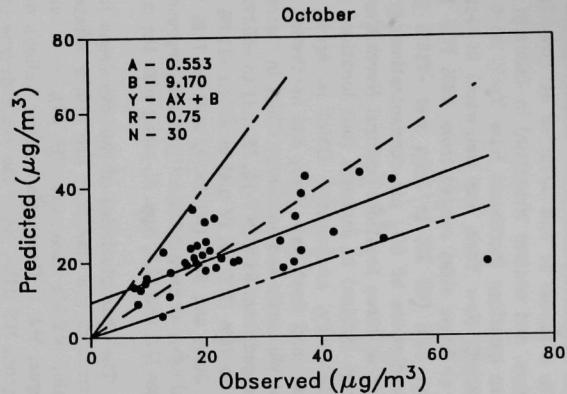
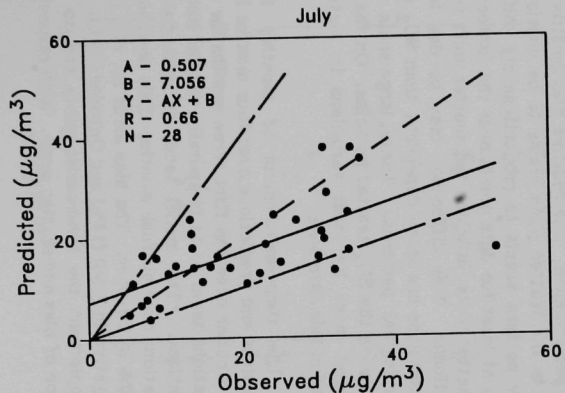
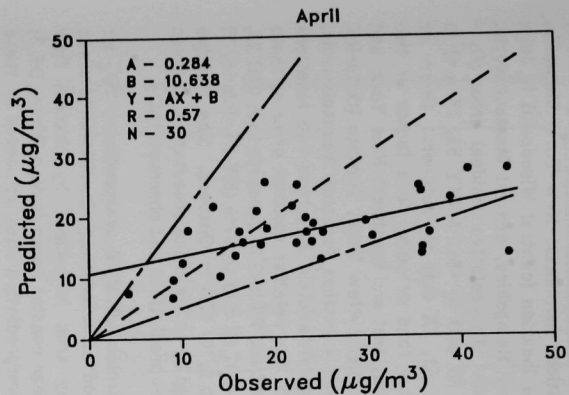
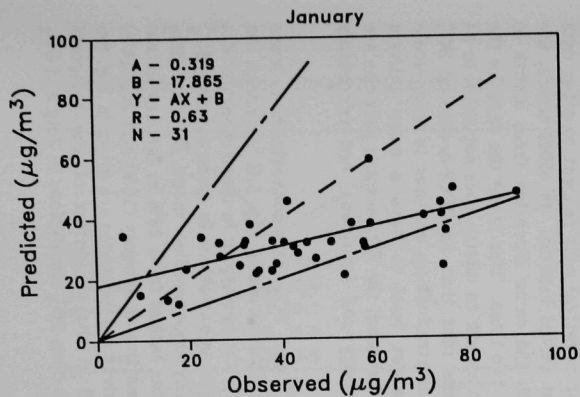


FIGURE 5.3 Monthly Scatter Plots of ASTRAP Predictions Versus Field Observations of Sulfate Air Concentrations in 1978



**FIGURE 5.4 Monthly Scatter Plots of ASTRAP Predictions Versus Field Observations of Sulfur Dioxide Air Concentrations in 1978**

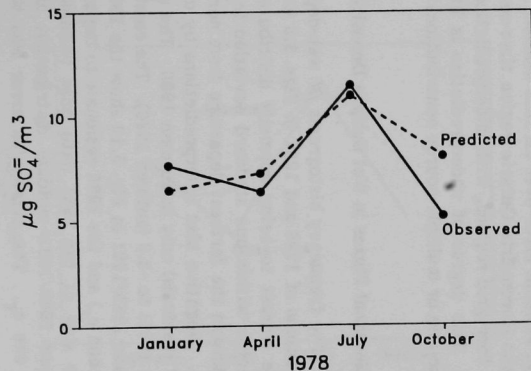
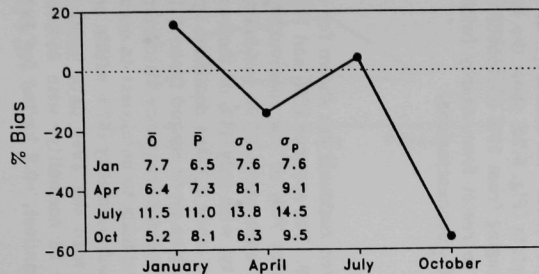
from the  $\text{SO}_2$  data plots. Over- or underpredictions of  $\text{SO}_4^-$  air concentrations greater or equal to a factor of two occurred at three locations (northeast New York, Memphis, Tennessee, and central Wisconsin) in January; one location (central Wisconsin) in July; and five locations (northeast New York, east New Hampshire, central Massachusetts, Long Island, New York, and Delaware) in October. The relative residual error ( $(O_i - P_i)/O_i$ ) at these sites ranged from 66% (TN 36, III b.1) to -510% (NY 51, IX a.4) in January; 54% (WI 39) in July; and -215% (NY 51, IX a.4) in October. Over- or underpredictions of  $\text{SO}_2$  air concentrations greater than or equal to a factor of two occurred at three locations (central North Carolina, east and southeast New York, and southwest Indiana) in January, four locations (southeast New York, Toronto (Ontario), west Kentucky, and central Ohio) in April; three locations (central Massachusetts, southeast and west New York, and northwest Pennsylvania) in July; and two locations (central Wisconsin and southeast Ohio) in October. The relative residual error at these sites ranged from 68% (NY 15, IX b.1) to -503% (NC 46, VII b.7) in January; 53% (OH 28, VI b.3) to 70% (NY 15, IV b.4) in April; -139% (MA 1, IX a.8) to 67% (NY 14, PA 16, VI c.3) in July; and 53% (WI 39, II a.7) to 71% (PA 2, VI d.7) in October. The relative residual error at each ASTRAP grid cell, represented by at least one observation, is given in Tables M.1 and M.2, App. M, along with the model predictions and observations.

The characteristics of the site areas for which ASTRAP overpredicted  $\text{SO}_4^-$  air concentrations are that they are either located on elevated/rough terrain (Whiteface Mountain, NY 51; Hanover, NH 37), in a valley area influenced by channel flow (Montague, MA 1; Connecticut River Valley), or near coastal areas (Indian River, DE 3; Huntington, NY 31). Since the same degree of overprediction does not occur at these sites for more than one month (except for Whiteface Mountain, two of three months with validation data), other factors besides complex terrain must be contributing to the poor model performance. The relative bias (positive or negative) with respect to other data points in the scatter plots seems to be consistent across months. In other words, the scatter pattern seems to proportionately shift in the same direction from month to month, at least for data points near the factor-of-two lines. This is not the case for the  $\text{SO}_2$  scatter plots, in which the scatter shift from month to month does not seem to be proportional. This difference may be due to the fact that ASTRAP-predicted  $\text{SO}_4^-$  concentrations are more dependent than  $\text{SO}_2$  concentrations on changes to large-scale meteorological patterns. These large-scale changes tend to have a more uniform influence on the  $\text{SO}_4^-$  scatter patterns. On the other hand, the  $\text{SO}_2$  scatter patterns are influenced more by local variations in meteorology and emissions, not adequately resolved in this data base.

The time-series plots of monthly  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentration mean observations and predictions over four months in 1978 is shown in Fig. 5.5. Although the bias is smaller over the first three months for  $\text{SO}_4^-$ , the variations in mean monthly  $\text{SO}_2$  observations seem to be temporally tracked better than variations in mean monthly  $\text{SO}_4^-$  observations. The model tends to systematically underpredict mean monthly  $\text{SO}_2$  observations over all four months. This positive bias ranges from 28% in January to less than 10% in October. The bias error for  $\text{SO}_4^-$  is positive in January (16%) and July (4%) and negative in April (14%) and October (54%). Univariate measures are given in Fig. 5.5 to complement the time-series plots and to aid in the evaluation of the temporal variation of bias and scatter error. With respect to bias,  $\text{SO}_4^-$  simulations in July and  $\text{SO}_2$



a.  $\text{SO}_4^-$  Bias Error and Mean Observations and Predictions



b.  $\text{SO}_2$  Bias Error and Mean Observations and Predictions

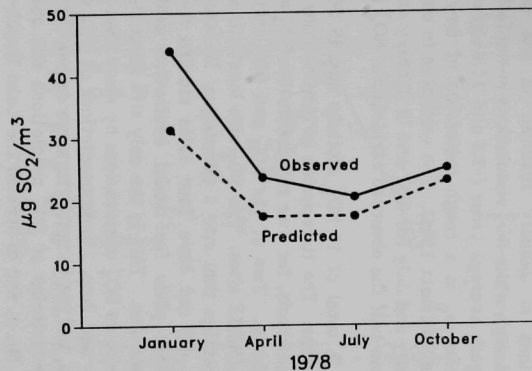
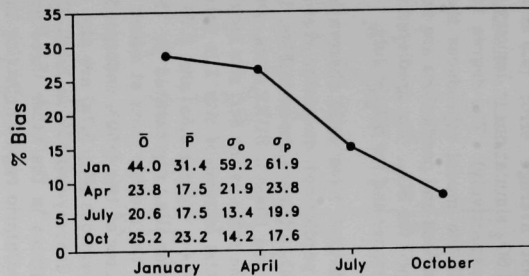


FIGURE 5.5 Time-Series Plots of Average Monthly Sulfate and Sulfur Dioxide Air Concentration Mean Observations and Predictions

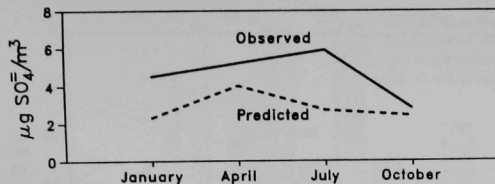
simulations in January resulted in the smallest average error (0.5 and  $3 \mu\text{g}/\text{m}^3$ , respectively), while  $\text{SO}_4^{2-}$  simulations in the fall and  $\text{SO}_2$  simulations in January resulted in the largest average error ( $-2.9$  and  $12.6 \mu\text{g}/\text{m}^3$ , respectively). The degree to which  $\sigma_p$  approaches  $\sigma_o$  is a rough indication of how well the model reproduces the observed variance (Wilmott 1984). The variances in observations and predictions are nearly equal in January and July for  $\text{SO}_4^{2-}$  and in January and April for  $\text{SO}_2$ . The predicted variance is about one-half the observed variance for  $\text{SO}_4^{2-}$  in October and for  $\text{SO}_2$  in July.

A total of 16  $\text{SO}_4^{2-}$  stations and 25  $\text{SO}_2$  stations have valid observations for all four months. The time series of observations and predictions, computed over a unit-grid increment basis, for four representative  $\text{SO}_4^{2-}$  and  $\text{SO}_2$  sites are shown in Figs. 5.6 and 5.7, respectively. Two of the  $\text{SO}_4^{2-}$  and  $\text{SO}_2$  stations are Class I, SURE sites, and two are Class II, SURE sites. The  $\text{SO}_2$  site locations are the same as the  $\text{SO}_4^{2-}$  site locations, with the exception that site 1 (Montague, Massachusetts) instead of site 3 is used for  $\text{SO}_2$ . (Site 3 did not have four valid months for  $\text{SO}_2$ .) Figures 5.6a and 5.7a show the time-series plots for model evaluation grid region IIa, represented by the Messer, Wisconsin site. This is the only site identified in the scatter plots at which the model underpredicts  $\text{SO}_4^{2-}$  observations by more than a factor of two ( $>50\%$  underprediction) for January and July. Underpredictions of  $\text{SO}_2$  observations (Fig. 5.7a) are also evident at this site, but not to quite the same degree ( $32\%$  in January and  $47\%$  in July). Comparison of time-histories of  $\text{SO}_4^{2-}$  predictions and observations in the other three grid regions (Figs. 5.6b, c, and d) showed the Lake Huron-Erie-Ontario region as having the smallest differences in January ( $14\%$ ) and April ( $-16\%$ ). The smallest differences in  $\text{SO}_4^{2-}$  observation-prediction error for July ( $-2\%$ ) and October ( $-42\%$ ) occurred in the southwest region at the Rockport SURE, Class I site (Fig. 5.6b). Similar comparisons of  $\text{SO}_2$  predictions with observations (Fig. 5.7b, c, d) showed the Rockport site to have the smallest residual error over all four seasons, ranging from  $<1\%$  (January) to  $12\%$  (April). The Lake Huron-Erie-Ontario region time-series plots (Fig. 5.7d) show the largest error of the four grid regions, with underpredictions ranging from  $20\%$  (October) to over  $65\%$  (July). This degree of underprediction is likely to result from nearby (within  $100 \text{ km}$ ) large-source or multiple-source contributions to  $\text{SO}_2$  concentrations.

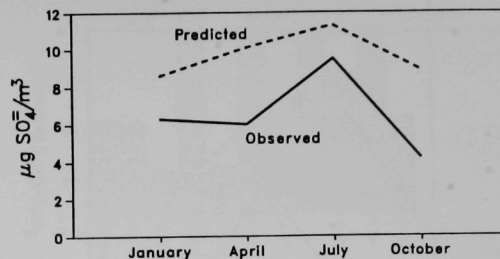
### 5.1.2 Seasonal Fluxes in Wet Sulfate Deposition

The frequency histograms of wet-deposition residuals are shown for each of the eight seasons of 1980 and 1981 in Figs. 5.8 and 5.9. The summer 1980 and 1981 residuals are the closest to being normally distributed, with  $70\%$  of the observation-prediction differences within one standard deviation ( $\sigma_p$ ) of each other. The residuals for winter 1980 showed the farthest departure from normality, with only  $44\%$  of the residuals within one  $\sigma_p$ . Negative bias (overpredictions by one or more  $\sigma_p$ ) was dominant (greater than  $50\%$  of residuals) only in autumn 1980. The mean bias error ranged from  $-0.1 \text{ kg SO}_4^{2-}/\text{ha}$  (spring 1981) to  $-3.0$  (autumn 1980). The combined season frequency histograms for 1980, 1981, and 1980/1981 in Fig. 5.10 show the 1981 residuals to be closest to normality ( $76\%$  within one  $\sigma_p$ ) and the 1980 residuals to depart from normality ( $47\%$  within one  $\sigma_p$ ). The greater size of the 1981 data base (compared with the 1980 data base) pushed the combined 1980/1981 residual distribution closer to normality, with  $65\%$  of residuals within one  $\sigma_p$ . The mean bias error was the smallest,  $-0.8 \text{ kg}/\text{ha}$ , for 1981, and the

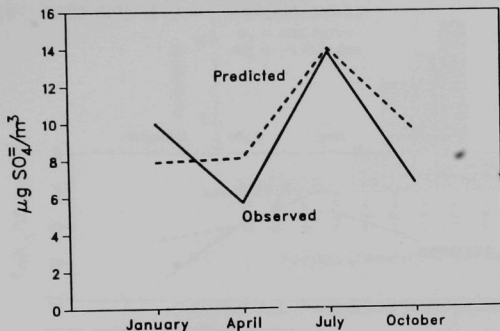
- a. SURE Class II Site; Messer, WI, #39  
Lake Michigan Region, II.a.7



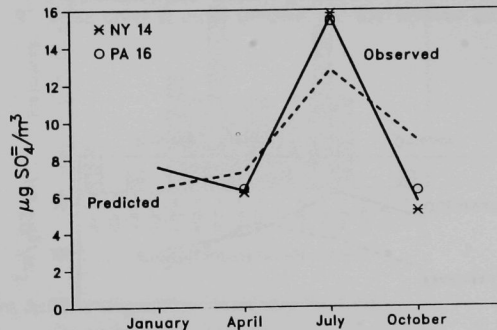
- c. SURE Class I Site; Indian River, DE, #3  
Atlantic Coast Region, X.a.1



- b. SURE Class I Site; Rockport, KY, #5  
Southwest Region, III.c.1

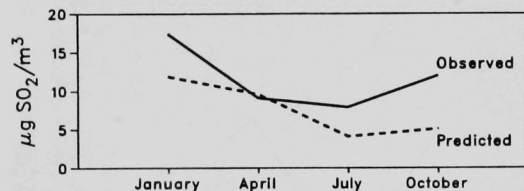


- d. SURE Class II Sites; Dunkirk, NY, #14; Warren, PA, #16;  
Lake Huron-Erie Ontario Region VI.c.3

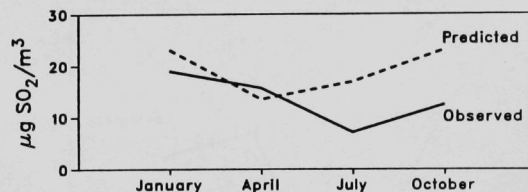


**FIGURE 5.6 Grid Region Time-Series Plots of Average Monthly Sulfate Air Concentration Observations and Predictions in Selected Subregions**

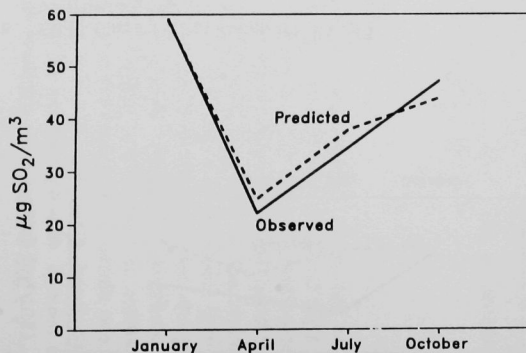
a. SURE Class II Site; Messer, WI, #39  
Lake Michigan Region, II.a.7



c. SURE Class I Site; Montague, MA, #1  
Northeast Region, IX.a.8



b. SURE Class I Site; Rockport, KY, #5  
Southwest Region, III.c.1



d. SURE Class II Sites; Dunkirk, NY, #14; Warren, PA, #16  
Lake Huron-Erie-Ontario Regions

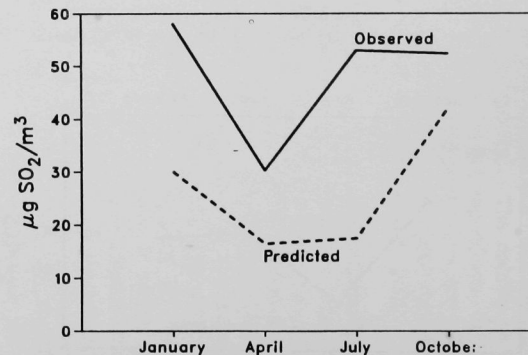
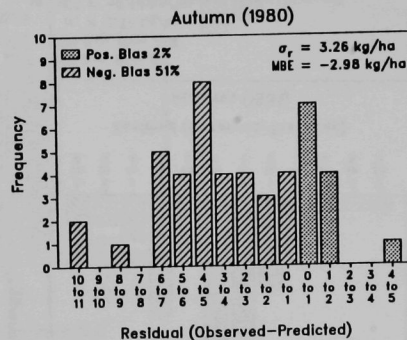
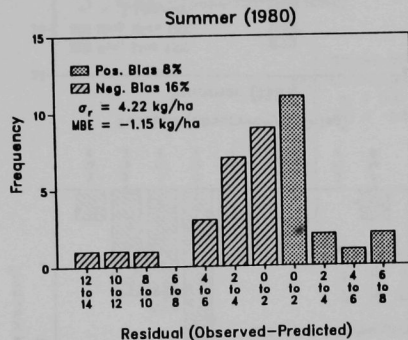
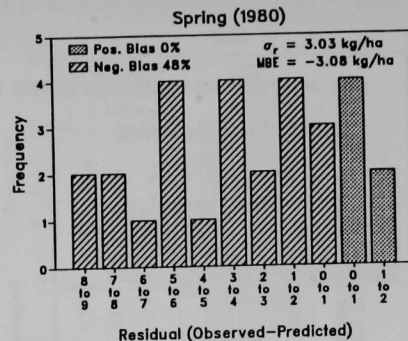
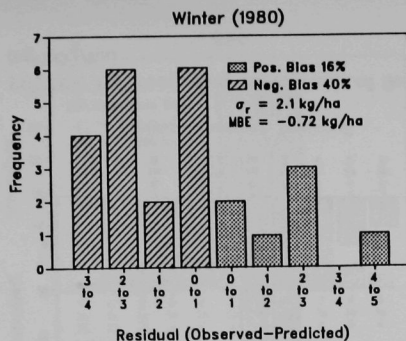
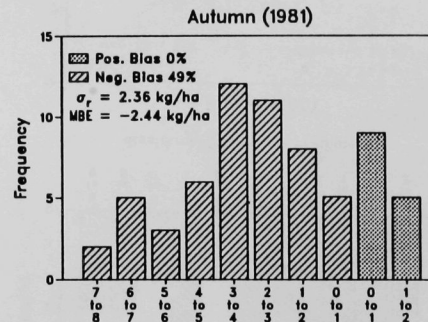
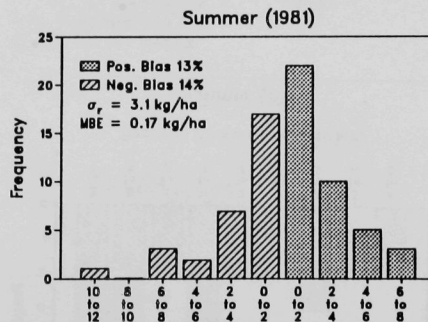
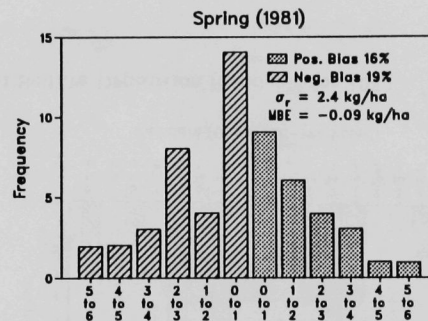
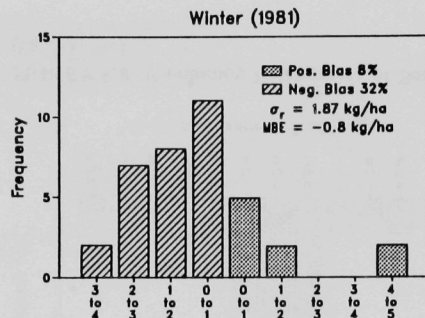


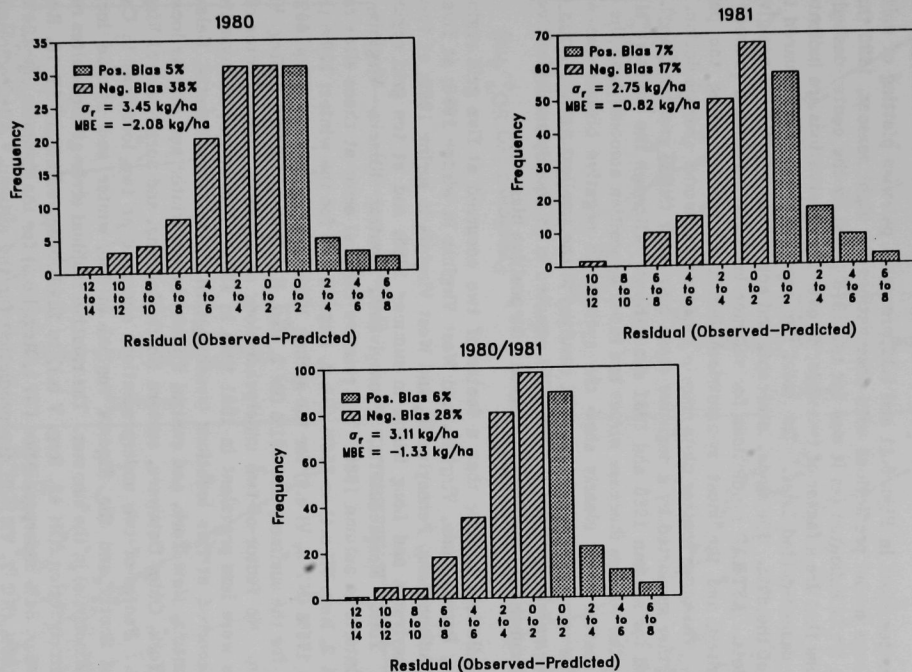
FIGURE 5.7 Grid Region Time-Series Plots of Average Monthly Sulfur Dioxide Air Concentration Observations and Predictions in Selected Subregions



**FIGURE 5.8 Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for 1980**  
(kg  $\text{SO}_4^-/\text{ha}$ )



**FIGURE 5.9 Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for 1981 ( $\text{kg SO}_4^{2-}/\text{ha}$ )**



**FIGURE 5.10** Frequency Histograms of Seasonal Wet Sulfate Deposition Residuals for Combined Seasons in 1980, 1981, and 1980/1981



largest, -2.01 kg/ha, for 1980. The mean bias errors and standard deviations for each season and the combined seasons are also given in the figures. The general shape of the observations and prediction distributions for combined like seasons and combined seasons for 1980, 1981, and 1980/1981 are shown in the frequency histograms in Figs. F.1 through F.3, App. F.

The scatter plots in Figs. 5.11 and 5.12 present a pairwise plotting of unit-grid increment average model predictions with observations for eight seasons, 1980 through 1981. Perfect fit is indicated (as it was for the 1978 plots) by the center dashed line, while predictions that are a factor of two under or over the observations are indicated by the two outer dashed-dotted lines. The solid line represents the least-squared linear regression fit of the data. The slope, intercept, and correlation coefficient are given in each of the plots. ASTRAP predictions for summer 1980 and spring and summer 1981 show the smallest and the most symmetrical (unbiased) scatter along the perfect prediction line. The other scatter plots show a tendency for model overprediction. This tendency is further supported by a negative mean bias error (MBE) greater than 2.0 for spring 1980 and for autumn 1980 and 1981 simulations. Although the negative MBE is small ( $<1.0$ ) for the winters (because winter has lower deposition amounts than the other seasons), the scatter plots clearly show the systematic negative bias in the winter predictions. The comparisons in spring 1981 resulted in the smallest negative bias (MBE = -0.1), while comparisons in the summer 1981 resulted in an equally small positive bias (MBE = 0.2). Summer 1981 was the only season with positive bias.

Overpredictions greater than a factor of two occurred at five grid regions in southwest Ohio, Pennsylvania, Virginia, and West Virginia in winter 1980; at five grid regions in the Adirondacks, Pennsylvania, and West Virginia in spring 1980; at two grid regions in Pennsylvania and Long Island in summer 1980; and at ten grid regions in western New York, Massachusetts, Pennsylvania, central Illinois, Virginia, and southeastern Ontario in autumn 1980. The relative residual error at these sites ranged from 100% (VA 2, Reg. VII c.4) to 256% (PA 3, Reg. VI d.7) for the winter; 129% (NY 3, Reg. VI c.6) to 175% (NY 6, VI c.9) for the spring; 103% (PA 2, Reg. VI d.5) to 166% (PA 3, Reg. VI d.7) for the summer; and 101% (IN 2, Reg. II d.7) to 387% (NY 3, Reg. VI c.6) for the autumn. No factor-of-two underpredictions occurred in 1980. Factor-of-two overpredictions were less prevalent in 1981 than in 1980. This degree of mismatching observations occurred at five locations (southwestern Ohio, Massachusetts, Delaware, Whiteface Mountain, New York, and central Illinois) in the winter and at nine locations (eastern New York, Ohio, Delaware, eastern North Carolina, and north central Virginia) in the autumn. Factor-of-two underpredictions occurred at two locations in Canada (northern Nova Scotia and the Algoma region) in the winter and at one location (northwestern Wisconsin) in the summer. The relative residual error at these sites ranged from 64% underprediction (OH 42, Reg. V b.1) to 214% overprediction (NY 10, Reg. IX a.4) for the winter; 64% underprediction (WI 1, Reg. I b.5) for the summer; and 104% (OH 4, VI b.2) to 262% (NC 7, VII c.9) overprediction for the autumn. All predictions for spring 1981 were within a factor of two of observations.

The factor-of-two overprediction data points in Figs. 5.11 and 5.12 predominantly represent precipitation chemistry samplers with event (MAP3S) or daily

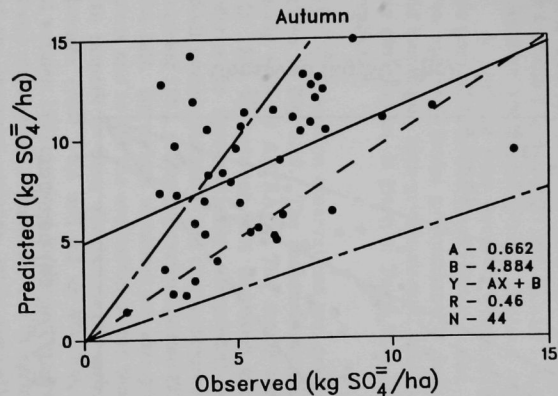
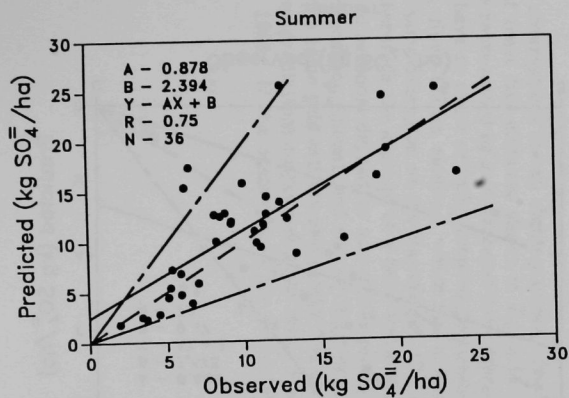
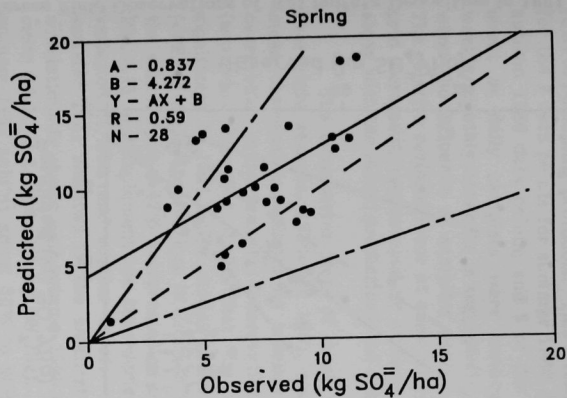
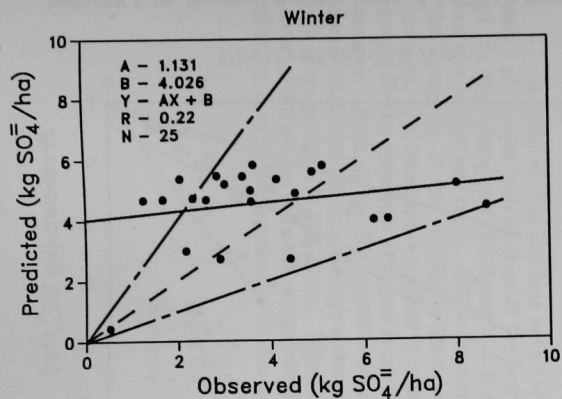


FIGURE 5.11 Seasonal Scatter Plots of ASTRAP Predictions Versus Field Observations of Wet Sulfate Deposition in 1980

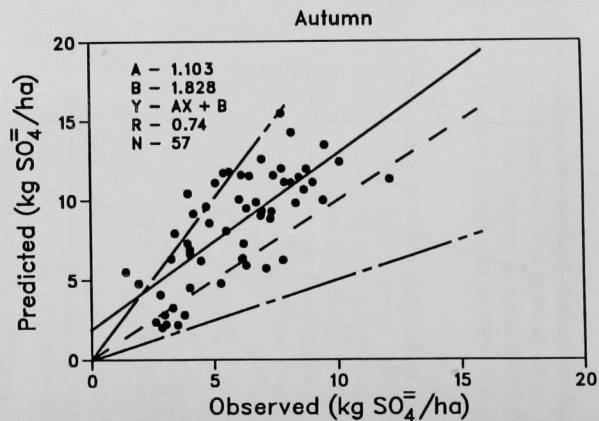
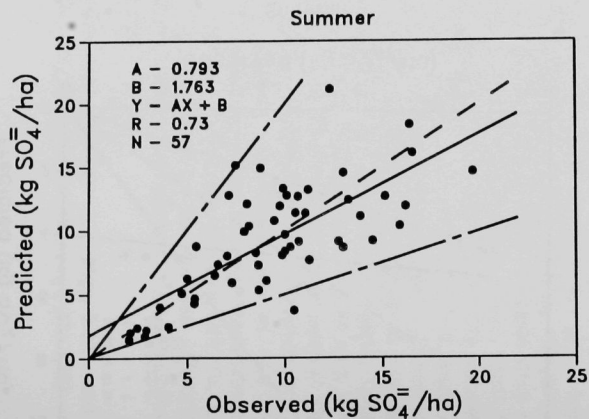
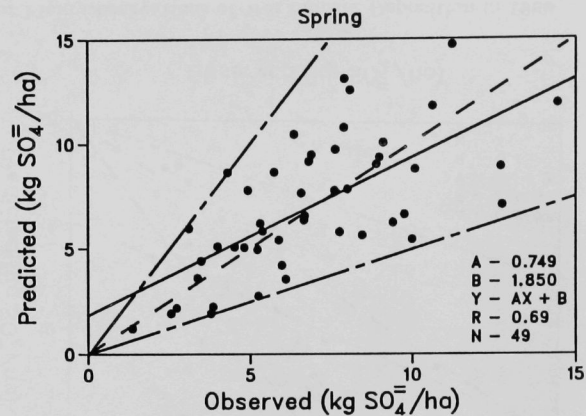
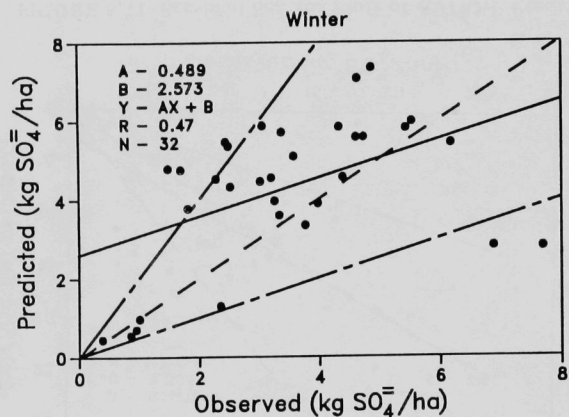


FIGURE 5.12 Seasonal Scatter Plots of ASTRAP Predictions Versus Field Observations of Wet Sulfate Deposition in 1981

(UAPSP) sampling protocols. This situation is true for all 5 data points for both winters, for the 2 data points for summer 1980, for 4 of the 5 spring 1980 data points, 7 of the 10 autumn 1980 data points, and 5 of the 9 autumn 1981 data points. Thirteen separate event or daily samplers were involved. A possible bias between event/daily and weekly/monthly networks is suggested. This possibility is addressed further in Sec. 5.4.4, when the influence of sampling protocol on apparent model performance is evaluated. The relative residual error at each unit-grid increment, nine-grid increment, and twelve-grid increment, represented by at least one observation, is given in Table M.3, App. M, along with the model predictions and observations.

The time-series plots of mean wet  $\text{SO}_4^-$  deposition observations and predictions over eight seasons, 1980 and 1981, are given in Fig. 5.13. The seasonal variations in mean observation seem to be tracked fairly well by ASTRAP. The model tends to overpredict mean seasonal observations in 1980 by between 12% (summer) and 54% (winter). The plot shows that the tendency to systematically overpredict is less significant in 1981. The model shows small overprediction (~1%) and underprediction (<4%) of mean observations in summer and spring 1981, moderate overprediction (23%) in the winter, and a 42% overprediction in the fall. Univariate measures are also given in Fig. 5.13 to complement the time-series plots and to aid the evaluation of the temporal variation of bias and scatter error. With respect to bias, simulations in spring and summer 1981 resulted in the smallest average error (-0.1 and 0.3 kg  $\text{SO}_4^-/\text{ha}$ ), while simulations in spring and autumn 1980 and autumn 1981 resulted in the largest average error (-3.1, -3.0, and -2.5 kg  $\text{SO}_4^-/\text{ha}$ ). Model simulations in summer, and in spring and autumn 1981 exhibit about 50% to 60% of the observed variance, while simulations in winter 1980 exhibit only 5% of the observed variance. The explained bias-corrected variance (EBCV) between seasons, decomposed into a spatial and temporal component, will be discussed in Sec. 5.3.2.

A total of five stations have valid observations for all eight quarters. The time series plots over individual grids for four of these five sites are shown in Fig. 5.14. Two of the four subregions (X.d.6 and VI.a.8) are represented by grid averages from more than one site in that subregion. All the stations have an event or daily sampling protocol, with the exception of the CANSAP Shelburne, Nova Scotia monitor, which has a monthly protocol. The Shelburne monitor is in the same subregion as the Kejimikujik APN daily sampler. The subregion seasonal mean is therefore the average of a daily and a monthly sampler. Each of the site-specific observed seasonal deposition amounts is plotted for comparison with the site average seasonal deposition amounts and the ASTRAP-predicted site average value in Fig. 5.14a. (Sites in the same grid cell are averaged.) The residual error or difference between observations and predictions ranges from less than 1% in fall 1981 to a 35% overprediction in summer 1980. If the observations are corrected for

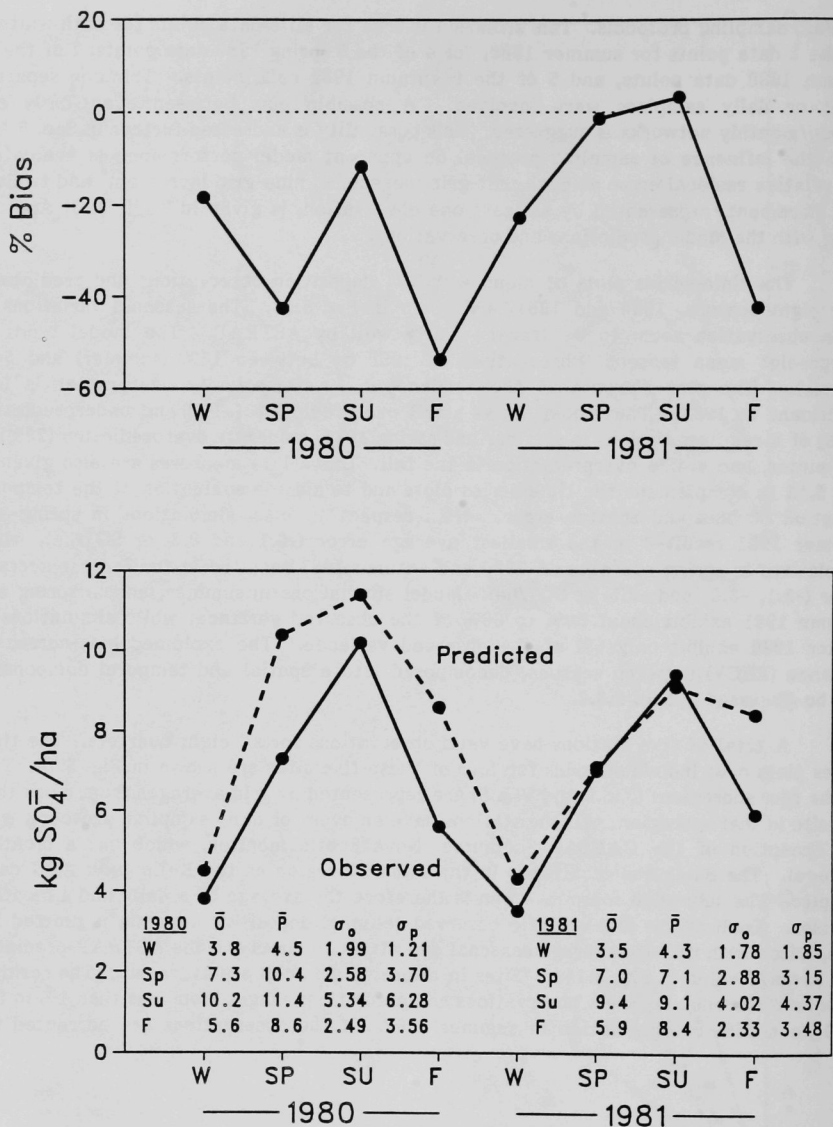
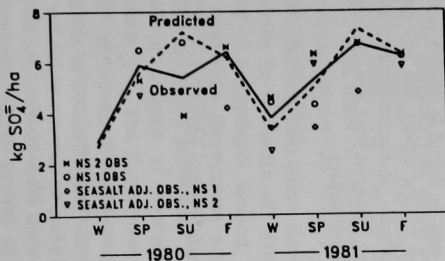
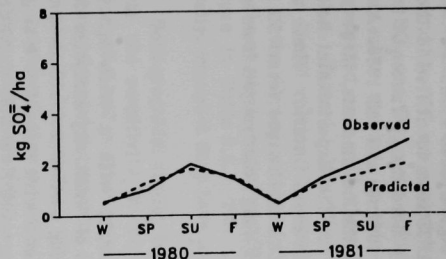


FIGURE 5.13 Time-Series Plots of Average Wet Sulfate Deposition Observations and Predictions

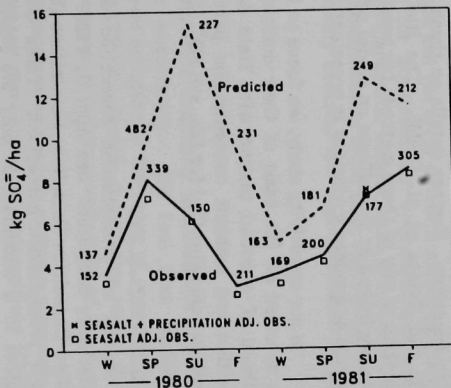
- a. Atlantic Coast Region X.1.6; Subregion X.d.6; Sites: NS2,1  
Kejimikujik, NS – APN, Daily; Shelburne, WS – CANSAP, Monthly



- c. Lake Superior Region Ia.1; Site: ON 49  
Ela, ONT – APN, Daily



- b. Northeast Region IX.b.7; Site: NY 1  
Brookhaven, NY – MAP3S, Event



- d. Lake Huron-Erie-Ontario Region VI.a.8; Sites: ON 10, 11, 12  
Long Point, ONT – APN, Daily; North East Hope, ONT –  
APIOS-D, Daily; Wellesly, ONT – APIOS-D, Daily

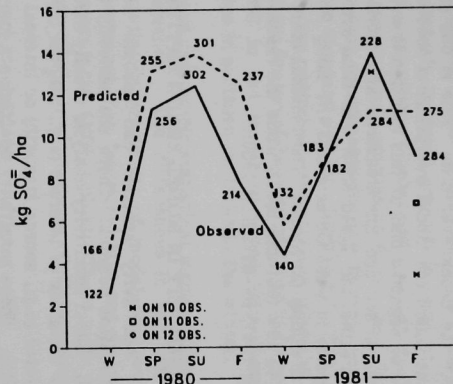


FIGURE 5.14 Time-Series Plots of Average Wet Sulfate Deposition Observations and Predictions in Selected Subregions

seasalt, the negative bias becomes slightly smaller.\* Figure 5.14b shows the time-series plot for the Brookhaven site. This is one of the MAP3S sites identified in the scatter plots at which the model overpredicts observations by more than a factor of two. This degree of negative bias at this location is evident for summer and fall 1980. Also shown for comparison are the measured and modeled precipitation amounts, in centimeters. Figures 5.14c and 5.14d show the time-series plots for the Experimental Lakes Area (ELA) site in west Ontario and the Long Point site in south Ontario. These two have daily sampling protocols. The residual error in the ELA site ranged from -3% (autumn 1980) to 29% (autumn 1981), while the residual error at Long Point ranged from about 1% (spring 1981) to -61% (fall 1980).

## 5.2 SENSITIVITY IN MODEL ERROR PATTERNS

Sensitivity analysis can assume several forms, with a wide spectrum of sophistication possible. There are basically two types of sensitivity analyses: local and global. Local sensitivity analysis usually involves taking partial derivatives of specified output variables of interest (e.g.,  $\text{SO}_4^-$  air concentrations) with respect to a single input parameter (e.g., source strength) or internal model parameter variable (e.g., deposition velocity). These sensitivity coefficients (partial derivatives) provide direct information about the effect that variations (small or large) in each parameter around its nominal value have on the output or state variables. Global sensitivity analysis involves taking partial derivatives of state variables with respect to all parameters simultaneously. The sensitivity coefficients in this case are local gradients of each variable with respect to each parameter in the multidimensional parameter space. Sensitivity information for local sensitivity analysis is obtained by Taylor series expansion or through solution of a set of ordinary partial differential equations. Sensitivity information for global sensitivity analysis is obtained by the Fourier Amplitude Sensitivity Test, pattern search procedures, or Monte Carlo method. These methods are used to determine probability density functions. Further details on these methods can be found in Tilden and Seinfeld (1982) and Rabitz et al. (1983). The limited scope of our study did not permit the application of any of these methods for sensitivity analysis. In our sensitivity evaluation, we simply looked at the effects on model performance patterns that are associated with individual or grouped variations in preselected model parameters.

In our sensitivity study, we focus on four internal model parameters. Parameters were selected on the basis of their commonality with the structure of other models, likelihood of eventual testing with field data, and general importance in influencing deposition and ambient concentration patterns. The parameters examined are dry-deposition velocities ( $V_d$ ) for  $\text{SO}_2$  and  $\text{SO}_4^-$ , linear transformation rate ( $T_L$ ) of  $\text{SO}_2$  to  $\text{SO}_4^-$ ,

---

\*No seasalt or precipitation correction was made to the observations that were statistically compared. See App. G, Table G.1, for the identification of sites that the Unified Deposition Data Base Committee (UDDBC 1985a, b) recommended for a seasalt and/or precipitation correction. Table G.2 shows that these corrections to the data resulted in a maximum of only 3% improvement in overall model performance for winter simulations.



and bulk sulfur wet-removal coefficient (WC). The diurnal and seasonal variations in  $V_d$  and  $T_r$  and the WC are scaled by a factor of two (100% larger and 50% smaller).<sup>\*</sup> The  $V_d$ s for  $SO_2$  and  $SO_4^{2-}$  were adjusted concurrently to avoid unrealistic relative  $V_d$  rates between sulfur species. In this study, all comparisons of observations with predictions are aggregated on a grid whose size is approximately equivalent to the model's spatial resolution. Table 5.1 identifies the parameter adjustments for the 27 model variations tested. Model version 10 is the nominal or standard version of ASTRAP. The seasonal ranges in the diurnal patterns of the maximum, minimum, and average values of  $V_d$  and  $T_r$  for the nominal version of ASTRAP and the factor-of-two adjustments to those ranges are given in Table 5.2. The nominal WC in the model does not vary diurnally or seasonally, except it does have a lower value in northern latitudes in the winter.

Both graphical display techniques and parametric statistical measures are used to evaluate the sensitivity in model performance. The differences in the patterns of displayed bias and scatter error and the relative sensitivity of model performance to variations in model parameters can be illustrated with fractional bias and scatter error (FBSE) plots and normalized bias and scatter error (NBSE) plots. These plots can also display relative (between model versions) estimates of goodness of fit between model predictions and observations.

As discussed previously (Sec. 4.2.1), FBSE plots are used principally for evaluating local-scale, short-term model performance.<sup>†</sup> In FBSE plots, FBE in model predictions is plotted on the x axis, while FSE in those predictions is plotted on the y axis. In NBSE plots, normalized scatter (NS) in model predictions is plotted on the y axis and computed as the ratio of the standard deviation of residuals to the square root of the product of the standard deviation of the observations and predictions. Normalized bias (NB) in model predictions is plotted on the x axis and computed as the ratio of the mean bias error (MBE) and the square root of the product of the standard deviation of the observations and predictions.

The sensitivity of model performance and the error patterns emerging from these plots are discussed in the next section for predicted air concentrations and wet-deposition fluxes. The patterns that emerge (parameter clustering in groups of three) are then used with index of agreement (IOA), dimensionless MSE, rank score index (RSI), and systematic/unsystematic MSE to quantitatively describe and explain sensitivity patterns in model performance.

---

<sup>\*</sup>The factor-of-two parameter adjustments were originally selected for the Bayesian probability analysis of source receptor uncertainty. We feel that factor-of-two variations in  $V_d$  and  $T_r$  represent a realistic estimate of the range of uncertainty of the nominal values of these parameters in ASTRAP. Section 2.4 of this report discusses these parameters further, including the seasonal and diurnal variations of the  $V_d$  and  $T_r$ .

<sup>†</sup>It should be noted that FE plots have previously been used to evaluate a model's ability to reproduce the upper end of the frequency distribution of observations (Cox et al. 1985a, 1985b).

**TABLE 5.1 Internal Model Parameter Adjustments Used for Model Performance Sensitivity Tests**

	Low Dry- Deposition Velocity			Normal Dry- Deposition Velocity			High Dry- Deposition Velocity		
	$V_d, T_r, WC^a$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$	$V_d, T_r, WC$
Transformation rate	Low	Normal	High	Low	Normal	High	Low	Normal	High
Low wet-removal coefficient	5 $L^b$ L L	4 L N L	6 L H L	2 N L L	1 N N L	3 N H L	8 H L L	7 H N L	9 H H L
Normal wet-removal coefficient	14 L L N	13 L N N	15 L H N	11 N L N	10 N N N	12 N H N	17 H L N	16 H N N	18 H H N
High wet-removal coefficient	23 L L H	22 L N H	24 L H H	20 N L H	19 N N H	21 N H H	26 H L H	25 H N H	27 H H H

<sup>a</sup>Parameters:  $V_d$  = dry-deposition velocity  
 $T_r$  = transformation rate  
 $WC$  = wet-removal coefficient

<sup>b</sup>Parameter Adjustments: H = High      H = 2 N  
N = Nominal      L = 0.5 N  
L = Low

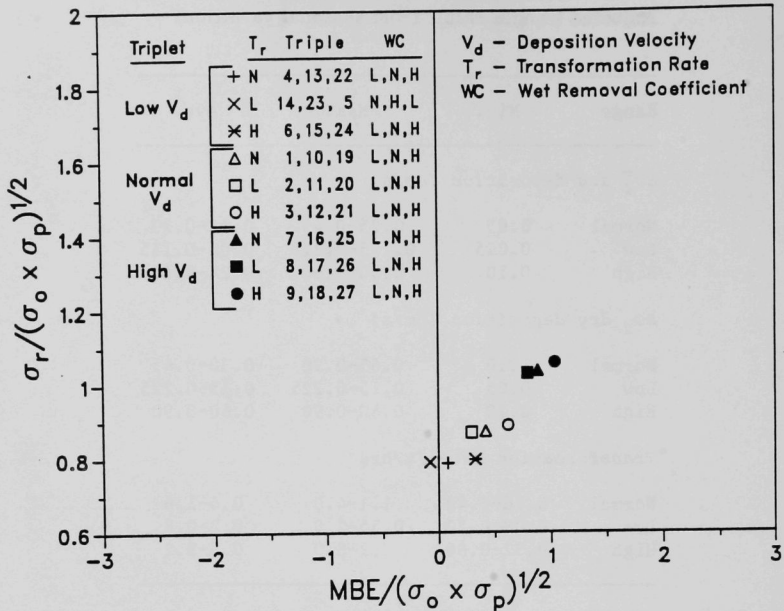
**TABLE 5.2 ASTRAP Dry-Deposition Velocity and Transformation Rate Variations, Normal and Adjusted (ranges result from seasonal variation)**

Range	Min.	Max.	Avg.
<i>SO<sub>4</sub><sup>=</sup> dry deposition (cm/s)</i>			
Normal	0.05	0.25-0.45	0.12-0.23
Low	0.025	0.125-0.225	0.06-0.115
High	0.10	0.50-0.90	0.24-0.46
<i>SO<sub>2</sub> dry deposition (cm/s)</i>			
Normal	0.10	0.65-0.90	0.30-0.45
Low	0.05	0.15-0.225	0.15-0.225
High	0.20	0.60-0.90	0.60-0.90
<i>Transformation rate (%/hr)</i>			
Normal	0.16-0.40	1.1-4.0	0.4-1.6
Low	0.08-0.20	0.55-2.0	0.2-0.8
High	0.32-0.80	2.2-8.0	0.8-3.2

### 5.2.1 Seasonal (Monthly Average) Air Concentrations and Monthly Fluxes in Wet Sulfate Deposition

The performance sensitivity patterns for ASTRAP simulations of July 1978 SO<sub>2</sub> air concentrations are shown in the NBSE plot of Fig. 5.15. The comparison of observations with all 27 versions of ASTRAP is represented. Each of the data points represents the approximate midpoint of the NB and NS of three parameter-set (PS) versions of ASTRAP (PS triplet). Each triplet is clustered in groups of three (triplets)\*, resulting from the factor-of-two adjustments to V<sub>d</sub>. The triplet with the smallest NS and NB error is the low V<sub>d</sub> triplet (V<sub>d</sub> held at one-half of its reference pattern of variation while varying T<sub>r</sub> and WC parameters) and is located in the lower center portion of the figure. The high V<sub>d</sub> triplet (V<sub>d</sub> held at twice the reference pattern of diurnal variation while adjusting the other parameters) has the largest NS and NB error and is located in the upper right of the figure. The normal V<sub>d</sub> triplet has NS and NB error between that of the other two triplets. Each triplet is ordered from left to right by low, normal, and high TR. If NB error is positive, parameter sets within each triplet are ordered from left to

\*A triplet is composed of model predictions from nine separate PS variations of ASTRAP, each of which is paired with the same set of observations.

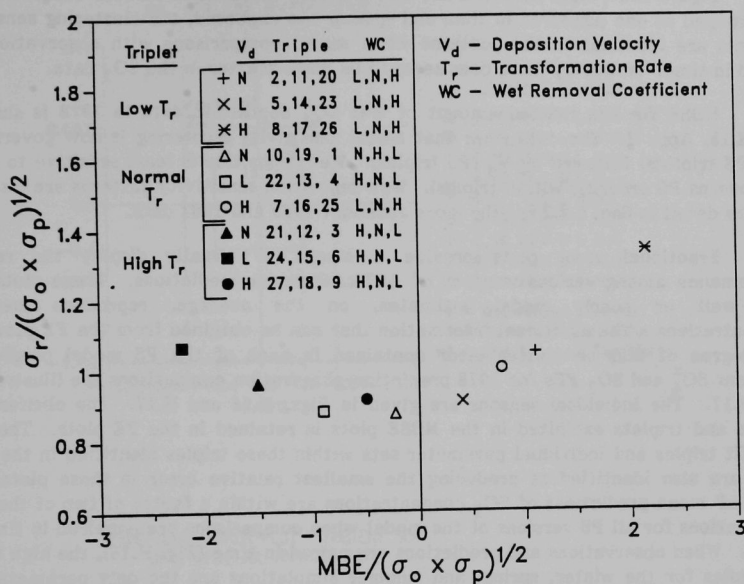


**FIGURE 5.15 Normalized Bias-Scatter Error Sensitivity Patterns for Sulfur Dioxide Air Concentrations**

right, by low, normal, and high WC. Negatively biased triplets have a reverse ordering: high, normal, and low WC. The low  $V_d$  triplet exhibits the smallest error (of the three triplets plotted) for the July 1978  $\text{SO}_2$  simulations. The low  $T_r$  triple within this triplet shows the best performance as measured by the smallest NB and NS error (which is proportionately equivalent to the smallest root mean same error).

Figure H.12, App. H, shows the NBSE plots of  $\text{SO}_2$  simulations for all four seasons (each data point represents the normalized bias and scatter of a single PS version of ASTRAP). The data show a positive bias tendency in ASTRAP predictions. This tendency is less significant in October and most significant in January. When  $V_d$  and  $T_r$  are high, positive bias and scatter error are the greatest. The bias and scatter error are minimized with nominal to low  $V_d$  and  $T_r$ , depending on the month being simulated.

Ambient  $\text{SO}_4^{2-}$  sensitivity patterns for ASTRAP parameter adjustments are given in Fig. 5.16. The data show the same clustering in triplets as in Fig. 5.15, but the clustering now results from the factor-of-two adjustments to  $T_r$  instead of  $V_d$ . Triplets are arranged from left to right by high, normal, and low  $T_r$ . The triplets within each triplet are ordered from left to right by low, normal, and high  $V_d$ . Parameter sets within



**FIGURE 5.16 Normalized Bias-Scatter Error Sensitivity Patterns for Sulfate Air Concentrations**

each triplet have the same ordering by WC as shown in the sensitivity pattern for  $SO_2$  air concentrations. The negatively biased triplets, as is the case for the  $SO_2$  patterns, have the reverse ordering of the positively biased triples. The normal  $T_r$  triplet exhibits the smallest error (of the three triplets plotted) for the July 1978  $SO_4^{2-}$  simulations. The normal  $V_d$  triplet within this triplet shows the best performance, as measured by the smallest NB and NS error (smallest MSE).

The NBSE plots of  $SO_4^{2-}$  simulations for all four seasons (each data point represents the NB and NS of a single PS version of ASTRAP) are shown in Fig. H.13, App. H. The data show a negative-bias tendency in ASTRAP predictions for the spring and fall and a slight positive-bias tendency in the winter. Comparing NBSE patterns for  $SO_4^{2-}$  and  $SO_2$  shows that ASTRAP simulations of  $SO_4^{2-}$  air concentrations are more sensitive to parameter variations than are ASTRAP simulations of  $SO_2$  air concentrations. The  $SO_4^{2-}$  plots clearly show that the clustering of model PS is governed in a hierarchical order by the output variables' (e.g., air concentrations of sulfate) sensitivity to model parameter variations. For  $SO_4^{2-}$ ,  $T_r$  variation has the greatest influence, followed by  $V_d$  and WC. For  $SO_2$ ,  $V_d$  variation has the greatest influence (but to a lesser degree than  $T_r$  has on  $SO_4^{2-}$  concentrations), followed by  $T_r$  and WC.

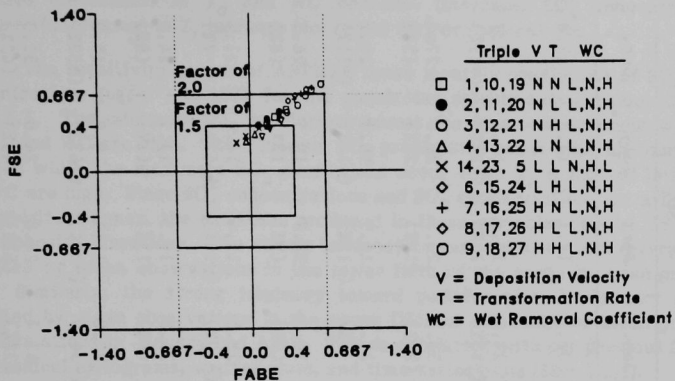
Figure H.14, App. H, shows the  $\text{SO}_2$  and  $\text{SO}_4^-$  NBSE when all four seasons of data are grouped as one (unpaired in time and space). As expected, the clustering sensitivity patterns are identical to the patterns when model comparisons with observations are paired in time. However, there does seem to be more scatter in the  $\text{SO}_2$  data.

NBSE for the limited amount of wet  $\text{SO}_4^-$  deposition data in 1978 is shown in Fig. H.15, App. H. The data show that model sensitivity clustering is now governed by WC (PS triplets) followed by  $V_d$  (PS triples). Wet deposition is least sensitive to the  $T_r$  ( $T_r$  governs PS ordering within triples). Wet-deposition sensitivity patterns are discussed in more detail in Sec. 5.2.2 for the more abundant 1980 and 1981 data.

Fractional error plots provide a means to visually display the relative performance among various versions of ASTRAP model predictions. These plots show how well or poorly model estimates, on the average, reproduce measured concentrations. The additional information that can be obtained from the FE plots is in the degree of bias or scatter error contained in each of the PS model predictions. Ambient  $\text{SO}_4^-$  and  $\text{SO}_2$  FEs for 1978 prediction-observation comparisons are illustrated in Fig. 5.17. The individual seasons are given in Figs. H.16 and H.17. The clustering by triples and triplets exhibited in the NBSE plots is retained in the FE plots. The same best-fit triples and individual parameter sets within these triples identified in the NBSE plots are also identified as producing the smallest relative error in these plots. The ASTRAP mean predictions of  $\text{SO}_2$  concentrations are within a factor of two of the mean observations for all PS versions of the model when comparisons are unpaired in time and space. When observations and predictions are paired in time (Fig. H.16), the high  $V_d$  and  $T_r$  triples for the winter, spring, and summer simulations are the only parameter sets that are projecting  $\text{SO}_2$  mean predictions greater than a factor of two of the mean observations. If all the fall simulations are within a factor of two of the mean observations, the scatter in mean  $\text{SO}_2$  predictions is within a factor of two of the scatter in mean  $\text{SO}_2$  observations for all PS versions. In the summer and fall, about half the parameter sets overpredict the mean scatter in observations in the winter and spring. Because of the greater sensitivity of  $\text{SO}_4^-$  air concentrations to PS variations (e.g.,  $T_r$ ), versions of ASTRAP overpredict mean observations of  $\text{SO}_4^-$  more than mean observations of  $\text{SO}_2$ . This is illustrated by the spread of the data in Figs. 5.17 and H.17. Figure H.18, App. H, shows the FSE and FAFE for the July 1978 wet  $\text{SO}_4^-$  comparisons.

The sensitivity in mean (over all receptors) ASTRAP simulations of air concentrations of monthly  $\text{SO}_4^-$  and  $\text{SO}_2$  to doubling and halving the internal model parameters individually, while holding the other model parameters to the nominal ASTRAP values, is given in Table 5.3. The mean predictions of the nominal or normal version of ASTRAP are also listed in the table for comparison. The numbers in parentheses are the absolute differences from the nominal ASTRAP predictions. As expected,  $\text{SO}_4^-$  air concentrations are most sensitive to variations in  $T_r$ , followed by variations in  $V_d$  and WC. Sulfur dioxide air concentrations are most sensitive to variations in  $V_d$ , followed by  $T_r$  or WC. Because no one PS strongly dominates sensitivity, as WC does for wet deposition and  $T_r$  does for ambient  $\text{SO}_4^-$ , the triple and triplet overlap in the NBSE and FSE plots (see Figs. 5.15, 5.17, and H.18) was greatest for  $\text{SO}_2$ . When all the parameters are doubled or halved simultaneously, these parameter adjustments show the greatest sensitivity (more than the individual variation of parameters) because  $\text{SO}_2$  concentration predictions are changed in the same direction.

FRACTIONAL ERROR PLOT - AMBIENT SO<sub>2</sub>  
JAN, APR, JUL AND OCT 1978



FRACTIONAL ERROR PLOT - AMBIENT SO<sub>4</sub>  
JAN, APR, JUL AND OCT 1978

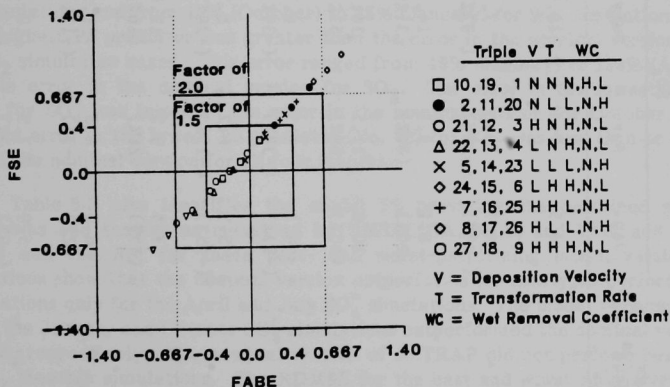


FIGURE 5.17 Fractional Error of Sulfur Dioxide and Sulfate Air Concentrations for Unpaired 1978 Predictions and Observations



**TABLE 5.3 Sensitivity in Model Mean Sulfate and Sulfur Dioxide Air Concentrations ( $\mu\text{g}/\text{m}^3$ ) to Variations in ASTRAP Internal Parameters**

Parameter Adjustments <sup>a</sup>	$\text{SO}_4^{=}$				$\text{SO}_2$			
	Jan.	Apr.	July	Oct.	Jan.	Apr.	July	Oct.
Nominal (10)	6.5	7.3	10.9	8.1	31.7	17.1	17.3	22.8
$2 \times V_d$ (16)	4.7 (-28%)	5.0 (-32%)	7.3 (-33%)	5.7 (-30%)	24.3 (-23%)	13.6 (-21%)	13.5 (-22%)	17.7 (-22%)
$0.5 \times V_d$ (13)	8.1 20%	9.5 (30%)	14.3 (31%)	10.3 (27%)	38.3 (21%)	20.3 (19%)	20.8 (20%)	27.2 (5%)
$2 \times T_r$ (12)	10.8 (40%)	11.5 (58%)	16.7 (53%)	13.1 (62%)	29.5 (-7%)	14.9 (-13%)	14.3 (-17%)	20.2 (-11%)
$0.5 \times T_r$ (11)	4.1 (-37%)	4.5 (-38%)	6.7 (-39%)	5.0 (-38%)	33.0 (4%)	18.7 (9%)	19.5 (13%)	24.5 (8%)
$2 \times \text{WC}$ (19)	5.5 (-15%)	6.2 (-15%)	8.7 (-20%)	7.1 (-12%)	28.5 (-10%)	15.4 (-10%)	14.4 (-17%)	21.3 (-7%)
$0.5 \times \text{WC}$ (1)	7.5 (15%)	9.2 (26%)	12.6 (16%)	9.4 (16%)	34.2 (8%)	19.3 (13%)	17.5 (1%)	24.6 (8%)
$2 \times \text{WC}, V_d, T_r$ (27)	6.6 (2%)	6.8 (-7%)	9.1 (-17%)	8.2 (1%)	20.8 (-34%)	11.1 (-35%)	9.8 (-43%)	15.2 (-33%)
$0.5 \times \text{WC}, V_d, T_r$ (5)	5.9 (-9%)	7.5 (3%)	10.7 (-2%)	7.4 (-9%)	43.7 (38%)	25.8 (51%)	24.8 (43%)	32.3 (42%)

<sup>a</sup> $V_d$  = dry-deposition rate;  $T_r$  = transformation rate; WC = wet-removal coefficient.

For  $\text{SO}_4^-$ , on the other hand, the effect of simultaneously doubling or halving all parameters produces the smallest changes in  $\text{SO}_4^-$  air concentrations. This is because increases (decreases) in  $V_d$  and WC decrease (increase)  $\text{SO}_4^-$  concentrations, while increases (decreases) in  $T_r$  increase (decrease)  $\text{SO}_4^-$  concentrations.

The sensitivity ranges of ASTRAP mean monthly predictions of  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations ( $\mu\text{g}/\text{m}^3$ ) in 1978 for the parameter adjustments considered are given in Table 5.4. The minimum  $\text{SO}_4^-$  predictions across the four seasons occur when  $T_r$  is low and  $V_d$  and WC are high. The minimum  $\text{SO}_2$  predictions occur when all four\* parameters are low, while the maximum  $\text{SO}_2$  predictions occur when  $T_r$  is high or low and both  $V_d$  and WC are high. Since  $\text{SO}_4^-$  concentrations and  $\text{SO}_2$  concentrations are affected by  $T_r$  in the opposite manner, the extremes produced in these variables require PS variations to be in opposite directions. The strong tendency toward negative bias (overpredictions) is exhibited by mean observations in the lower fifth of the range in mean predictions for  $\text{SO}_4^-$ . Similarly, the strong tendency toward positive bias, in January and April, is exhibited by mean observations in the upper fifth of the range of mean predictions for  $\text{SO}_2$  simulations in January and April. This is consistent with our previous findings made with residual histograms, scatter plots, and time-series plots (Sec. 5.1.1).

The relative dimensionless mean square error (RDMSE) or relative mean error, expressed as a percentage of the mean square observation (MSO), is a fairly good indicator of the percent overall bias and scatter error in model predictions. Table 5.5 gives the calculated RDMSE for the nominal version of ASTRAP and the parameter adjusted versions producing the upper and lower prediction extremes just identified. The relative mean error for the nominal version ranged from 4% (July) to 33% (October) for  $\text{SO}_4^-$  simulations and from 18% (October) to 25% (January) for  $\text{SO}_2$  simulations. The error in the highest PS predictor was greater than the error in the nominal version for all  $\text{SO}_4^-$  and  $\text{SO}_2$  simulation cases. This error ranged from 49% (January) to 124% (April) greater than the error in the nominal version for  $\text{SO}_2$ . The error in the lowest PS predictor (No. 6) for  $\text{SO}_4^-$  was less than the error in the nominal version for October simulations, while the error in the lowest PS predictor (No. 26) for  $\text{SO}_2$  was less than or equal to the error in the nominal version for all four months.

Table 5.5 also identifies the model PS providing the combined smallest and largest bias and scatter as measured by DMSE, IOA, MLE, and VLE, and it gives the RDMSE and the RSI for these best- and worst-performing model versions. These calculations show that the nominal version outperforms (or nearly outperforms) all other PS variations only for the April and July  $\text{SO}_4^-$  simulations. The model versions performing best in the January and October  $\text{SO}_4^-$  simulations outperformed the nominal version by 4% and 29%, respectively. The nominal version of ASTRAP did not perform best for any of the  $\text{SO}_2$  monthly simulations. The RDMSE for the best and worst PS performers ranged from 4% (PS 18 in July and PS 23 in October) to 209% (PS 6 in October) for  $\text{SO}_4^-$  simulations and from 12% (PS 4 in January and April) to 78% (PS 27 in July) for  $\text{SO}_2$  simulations. These ranges in error for the PS variations tested show that ASTRAP performance sensitivity is much greater for  $\text{SO}_4^-$  simulations than for  $\text{SO}_2$  simulations.

\* $V_d$  for  $\text{SO}_2$  and for  $\text{SO}_4^-$  are adjusted in the same direction simultaneously; thus,  $V_d$  variation can be considered a two-parameter variation.

TABLE 5.4 Sensitivity Ranges in ASTRAP for Four Seasons of Simulations ( $\mu\text{g}/\text{m}^3$ )

Perform. Measure <sup>a</sup>	$\text{SO}_4^=$				$\text{SO}_2$			
	Jan.	Apr.	July	Oct.	Jan.	Apr.	July	Oct.
$\bar{P}_i$	2.6-15.8	2.6-18.5	3.6-25.0	3.1-19.4	20.8-43.7	11.1-25.8	9.8-24.8	15.2-32.3
$T_r$	L - H	L - H	L - H	L - H	H/L - L	H/L - L	H/L - L	H/L - L
$V_d$	H - L	H - L	H - L	H - L	L - H	L - H	L - H	L - H
WC	H - L	H - L	H - L	H - L	L - H	L - H	L - H	L - H
PS	26 6	26 6	26 6	26 6	6/5 27	6/5 27	6/5 27	7/6 27
$\bar{O}$	7.7	6.4	11.1	5.3	43.4	22.8	20.3	24.7

<sup>a</sup> $\bar{P}_i$  = mean prediction over grid cell i  
 $T_r$  = transformation rate  
 $V_d$  = dry-deposition rate  
 WC = wet-removal coefficient  
 PS = parameter set  
 $\bar{O}$  = mean observation over all grid cells.

**TABLE 5.5 Relative Mean Square Error (%) in Sulfate and Sulfur Dioxide Air Concentration Predictions**

Performance	PS <sup>a</sup>	SO <sub>4</sub> <sup>=</sup>				PS	SO <sub>2</sub>			
		Jan.	Apr.	July	Oct.		Jan.	Apr.	July	Oct.
Lowest	26	140	86	144	25	5	12	12	16	18
Nominal	10	12	7	4	33	10	25	24	19	18
Highest	6	61	131	72	209	27	75	74	78	47
Best overall performance		8(18)	5(19)	4(18)	4(23)		12(4)	12(4)	15(13)	15(6)
RSI best		1.839	1.605	1.308	1.354		1.828	1.803	1.761	1.790
Worst overall performance		61(6)	131(6)	144(26)	209(6)		75(27)	74(27)	78(27)	47(27)
RSI <sup>b</sup> worst		5.229 (6)	11.078	5.127 (6)	11.174		3.739	3.688	5.320	2.889
		5.219 (26)		5.388 (26)						

<sup>a</sup>PS = parameter set.

<sup>b</sup>RSI = rank score index.

The highest predictor, PS 6 for  $\text{SO}_4^-$ , performed the worst for three of the four months. The lowest predictor, PS 26, performed the worst only for July  $\text{SO}_4^-$  simulations. The highest predictor, PS 27, performed worst for all four  $\text{SO}_2$  simulation months. Best  $\text{SO}_4^-$  performance came with high  $V_d$  and  $T_r$  and low WC (PS 18) in January and July, normal  $V_d$  and  $T_r$  and high WC (PS 19) in April, and low  $V_d$  and  $T_r$  and high WC (PS 23) in October. Best  $\text{SO}_2$  performance came with low  $V_d$  and WC and normal  $T_r$  (PS 4) in January and April, low  $V_d$  and normal WC and  $T_r$  (PS 13) in July, and low  $V_d$  and WC and high  $T_r$  (PS 6) in October.

The four performance measures of bias error and scatter error used to compute RSI for the nominal version of ASTRAP and the best- and worst-performance versions (Table 5.5) are given in Table 5.6. These measures identify ASTRAP's  $\text{SO}_4^-$  performance as being best for the summer and worst for the fall, and ASTRAP's  $\text{SO}_2$  performance as being best for the fall and summer and worst for the winter and spring. By decomposing mean square error into its systematic and unsystematic components,\* we can compute the minimum systematic MSE achievable through the factor-of-two adjustments to model parameters. Table 5.6 gives the percent MSE in ASTRAP predictions that results from systematic and unsystematic causes. Also given are the computed MSE and its systematic and unsystematic parts, the minimum percent of systematic MSE achievable with PS variation, the PS producing this minimum, the MSE components for this PS, and the systematic error reduction potential (SERP) achievable with the PS adjustments we used. The percent error that is systematic is small only for summer  $\text{SO}_4^-$ . More than 70% of the apparent model error is systematic for winter  $\text{SO}_4^-$  and  $\text{SO}_2$ , spring  $\text{SO}_2$ , and fall  $\text{SO}_4^-$  predictions. The SERP is greatest (59%) for model predictions of fall  $\text{SO}_4^-$ . Almost 100% of the systematic error in summer  $\text{SO}_4^-$  can be removed by PS adjustment. The remainder of the error is primarily unsystematic, which may indicate that not much model performance improvement can be achieved for July, short of model reformation or improving the spatial and temporal restriction of the meteorology, emissions, and net-deposition sampling data base.

### 5.2.2 Seasonal Fluxes in Wet Sulfate Deposition

The performance sensitivity patterns for ASTRAP simulations of summer 1980 wet  $\text{SO}_4^-$  deposition are shown in the NBSE plot of Fig. 5.18. Comparisons of observations

---

\*Systematic error is the error in the empirically derived data internal to the model (e.g., model PS), the model input to the data (e.g., to generate emission, wind, and precipitation fields), and the model evaluation data observed (e.g.,  $\text{SO}_4^-$  and  $\text{SO}_2$  concentrations and wet  $\text{SO}_4^-$  deposition). Systematic error is assumed to be reducible error. Refer to Table 1.1 in Sec. 1 of this report for our definition of reducible error. If the systematic error can be eliminated entirely, the remaining error (unsystematic error) can be interpreted as the potential accuracy of the model. Error reduction that would require model reformulation is considered in this report to be unsystematic error. Strictly speaking, since our sensitivity study was restricted to variations only in internal model parameters, the only reducible error available for our consideration was the systematic error inherent to the four model parameter sets.

**TABLE 5.6 ASTRAP Sulfate and Sulfur Dioxide Air Concentration Seasonal (monthly average) Performance and Systematic Error Reduction with Parameter Variation**

Performance Measure <sup>a</sup>	Winter		Spring		Summer		Fall	
	SO <sub>4</sub> <sup>=</sup>	SO <sub>2</sub>	SO <sub>4</sub> <sup>=</sup>	SO <sub>2</sub>	SO <sub>4</sub> <sup>=</sup>	SO <sub>2</sub>	SO <sub>4</sub> <sup>=</sup>	SO <sub>2</sub>
IOA	0.60	0.62	0.56	0.61	0.83	0.75	0.45	0.72
VLE	0.229	0.232	0.058	0.163	0.056	0.198	0.076	0.158
DMSE	0.132	0.309	0.690	0.294	0.044	0.254	0.251	0.225
MLE	0.105	0.254	-0.118	0.240	0.059	0.109	-0.429	0.019
RSI	2.13	2.41	2.65	2.35	1.36	1.90	2.97	1.80
% MSE <sub>u</sub>	27.9	14.3	73.3	16.2	91.6	47.0	23.9	39.1
% MSE <sub>s</sub>	72.1	85.7	26.7	83.8	8.4	53.0	76.1	60.9
MSE	6.6	427.5	3.3	122.3	5.5	91.5	10.7	131.1
MSE <sub>u</sub>	1.9	366.5	2.4	19.8	5.0	43.0	2.6	51.2
MSE <sub>s</sub>	4.8	61.0	0.9	102.5	0.5	48.5	8.1	79.9
Minimum % MSE <sub>s</sub>	48.4	63.5	5.1	62.0	0.2	33.9	17.5	50.9
PS	18	5	19	5	18	4	25	23
MSE	2.30	270.7	1.9	87.6	5.9	86.8	1.12	134.2
MSE <sub>u</sub>	1.65	171.1	1.8	36.1	5.9	57.4	1.02	65.9
MSE <sub>s</sub>	1.60	99.0	0.1	51.5	0.0	29.4	0.47	68.3
SERP <sup>b</sup>	24%	22%	22%	22%	8.2%	19%	59%	10%

<sup>a</sup>IOA = index of agreement  
VLE = variance logarithmic error  
DMSE = dimensionless mean square error  
MLE = mean logarithmic error  
RSI = rank score index  
MSE = mean square error  
MSE<sub>u</sub> = mean square error unsystematic  
MSE<sub>s</sub> = mean square error systematic  
PS = parameter set

<sup>b</sup>Systematic error reduction potential is approximate.

with all 27 versions of ASTRAP are represented. Each of the data points represents the approximate midpoint of the NB and NS of three PS versions of ASTRAP (PS triple). As for  $\text{SO}_4^-$  and  $\text{SO}_2$ , each triple is clustered in groups of three parameter sets (triplets)\* resulting from the factor-of-two adjustments to WC. (Triplets are arranged from left to right by high, normal, and low WC.) The triples within each triplet are ordered from left to right by  $V_d$ . Parameter sets within each triple are now ordered by the  $T_r$  in the same way that WC ordered individual parameter sets for ambient  $\text{SO}_2$  and  $\text{SO}_4^-$ . The distance of the data points from the origin in the NSBE plots is proportionately equivalent to the RMSE. The normal WC triplet exhibits the smallest error (of the three triplets plotted) for the summer simulations. The high-deposition-velocity triple within this triplet shows the best overall performance, as measured by the smallest NB and NS error. This triple (PS: 17, 16, 18) also has the smallest RMSE (3.8 kg  $\text{SO}_4^-/\text{ha}$ ) of the nine PS triples evaluated.

Figures H.19 and H.20, App. H, show the NBSE plots of wet  $\text{SO}_4^-$  deposition simulations for each of the eight seasons in 1980 and 1981. The data show a negative

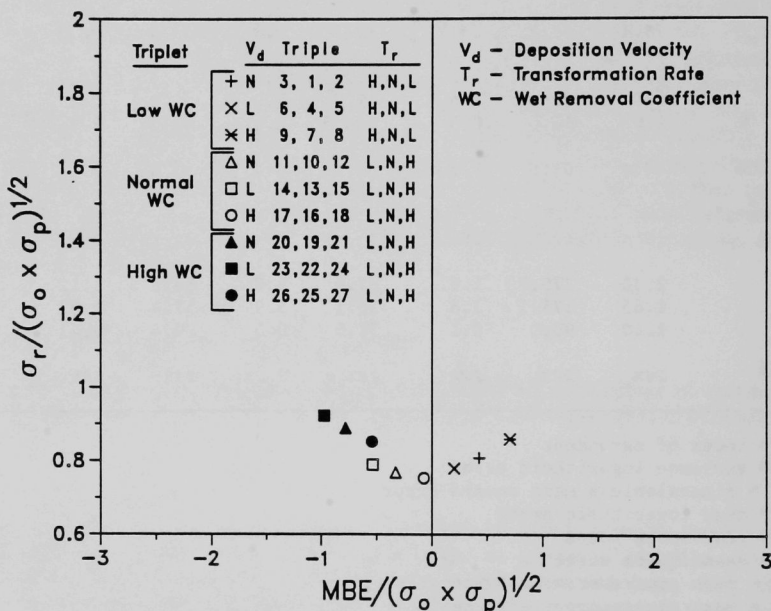


FIGURE 5.18 Wet-Deposition Sensitivity Clusters for Summer 1980

\*A triplet is composed of model predictions from nine separate PS variations of ASTRAP, each of which is paired with the same set of observations.



bias tendency in the winter and fall of both years and the spring of 1980. A much smaller degree of bias, particularly for the normal WC and  $V_d$  triple, is exhibited in the plots for both summers and for spring 1981. The NBSE patterns for wet deposition show that ASTRAP simulations of wet  $\text{SO}_4^-$  deposition are not as sensitive to parameter variations as ASTRAP simulations of  $\text{SO}_4^-$  air concentrations. However, the ordering or positioning of PS within each triplet for wet deposition is more sensitive to  $V_d$  than is the case for  $\text{SO}_4^-$  air concentrations. Figures H.21 and H.22 show the seasonal NBSE. The best-performing PS triples from the figures\* are 12, 10, and 11 (DMSE = 0.13) for summer; 16, 18, and 17 (DMSE = 0.25) for winter; 17, 16, and 18 (DMSE = 0.14) for spring; and 3, 1, and 1 (DMSE = 0.11) for autumn. Figure H.23 shows the NBSE for 1980 and 1981, unpaired with four season groupings. PS triple 17, 16, 18 (DMSE) performs best for both 1980 and 1981 data, with DMSE = 0.206 for 1980 and DMSE = 0.147 for 1981. In this case, if the minimum DMSE is used exclusively to rank performance, PS triple 6, 4, 5 (DMSE = 0.183) would be ranked as performing best for 1980. If the index of agreement (IOA) is used, PS triple 17, 16, 18 would just barely outperform PS triple 6, 4, 5. Further examination of the NBSE plots shows that the scatter error is smallest for PS 17, 16, 18 (which is ranked best by IOA) and the bias error is smallest for PS 6, 4, 5 (which is ranked best by DMSE). Therefore, DMSE tends to favor the smallest bias error while IOA tends to favor the smallest scatter error.<sup>‡</sup>

Fraction error plots for the winters and summers of 1980 and 1981 are shown in Fig. 5.19. The same best-fit triples identified in the NBSE plots are also identified as producing the smallest relative error in these plots. Figures 5.19c and d (summers) show the ASTRAP mean predictions of wet deposition are within a factor of two of the mean observations for all PS versions of the model (observations and predictions paired in time). Spring 1981 (Fig. H.24) is the only other season in which the predictions from all PS versions of ASTRAP were within a factor of two of observations. The scatter in the predictions with the low WC triplet (nine PS) and the bias in the prediction with the high WC and low  $V_d$  triple (PS 22, 23, 24) are a factor of two greater than the winter 1980 observations (Fig. 5.19a). Both the scatter and bias in the predictions with the high WC and low  $V_d$  triple (PS 22, 23, 24) are a factor of two greater than winter 1981 observations (Fig. 5.19b). The parameter sets for other seasons (autumn 1980 and 1981 and spring 1980) with predictions not within a factor of two of observations are identified in Fig. H.24a, c, and d.

The sensitivity of ASTRAP simulations of seasonal wet  $\text{SO}_4^-$  deposition to doubling and halving the internal model parameters individually, while holding the other model parameters at the nominal ASTRAP rates, is given in Table 5.7. The nominal

---

\*These triples are also identified as performing best for the above seasons when the DMSE is used exclusively as the performance measure.

<sup>‡</sup>Because of the tendencies of DMSE and IOA toward the smallest bias or smallest scatter respectively, a combined measure was developed (RSI) using DMSE, IOA, MLE, and VLE.

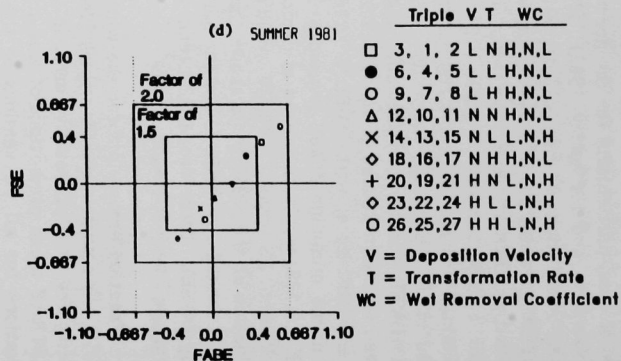
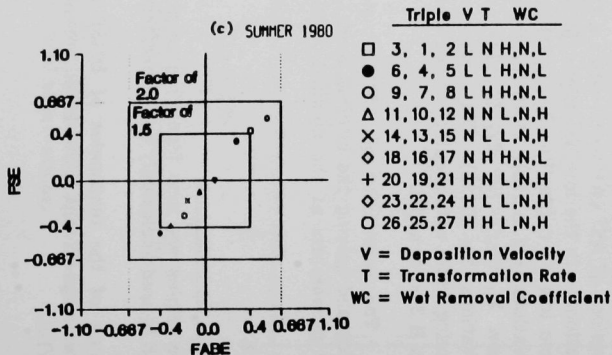
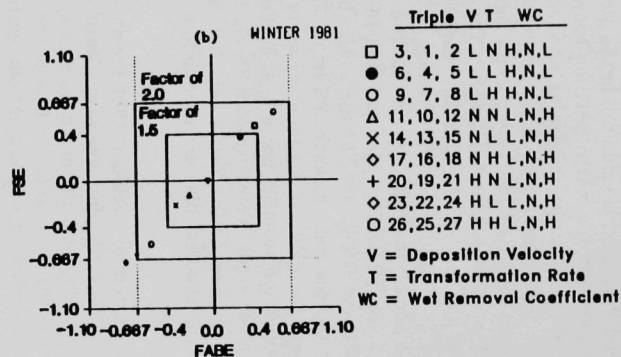
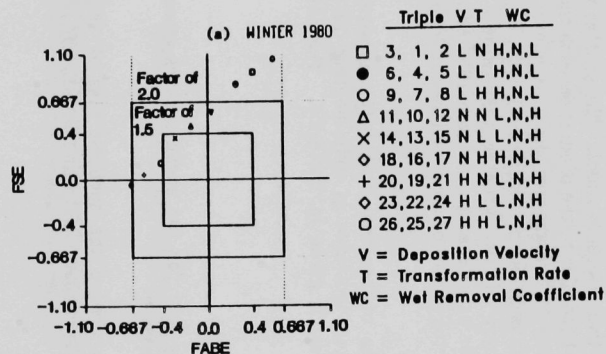


FIGURE 5.19 Fractional Error Sensitivity for Winter and Summer, 1980 and 1981

**TABLE 5.7 Sensitivity in Model Mean Wet Sulfate Deposition (kg/ha) to Variations in ASTRAP Internal Parameters**

Parameter Adjustments <sup>a</sup>	1980				1981			
	W	Sp	Su	F	W	Sp	Su	F
Nominal (10)	4.5	10.4	11.4	8.6	4.3	7.1	9.2	8.4
2 x WC (19)	6.8 (+51%)	13.2 (+27%)	14.4 (+26%)	11.3 (+31%)	6.7 (+56%)	8.8 (+24%)	11.3 (+23%)	10.1 (+20%)
0.5 x WC (1)	2.8 (-39%)	7.0 (-33%)	7.5 (-34%)	5.6 (-35%)	2.6 (-40%)	4.8 (-32%)	6.3 (-32%)	5.8 (-31%)
2 x V <sub>d</sub> (16)	3.9 (-13%)	9.1 (-13%)	10.1 (-10%)	7.5 (-13%)	3.7 (-14%)	6.2 (-13%)	8.0 (-13%)	7.3 (-13%)
0.5 x V <sub>d</sub> (13)	5.2 (+16%)	11.5 (+11%)	12.6 (+11%)	9.6 (+12%)	4.7 (+9%)	7.9 (+11%)	10.3 (+12%)	9.3 (+11%)
2 x T <sub>r</sub> (12)	4.6 (+2%)	10.5 (+1%)	11.5 (+1%)	8.7 (+1%)	4.35 (+1%)	7.2 (+1%)	9.4 (+2%)	8.5 (+1%)
0.5 x T <sub>r</sub> (11)	4.45 (-1%)	10.3 (-1%)	11.3 (-1%)	8.55 (-1%)	4.25 (-1%)	7.0 (-1%)	9.1 (-1%)	8.3 (-1%)
22 x WC, V <sub>d</sub> , T <sub>r</sub> (27)	5.9 (+31%)	11.9 (+14%)	13.1 (+15%)	10.1 (+17%)	5.9 (+37%)	7.9 (+11%)	10.1 (+10%)	9.0 (+7%)
0.5 x WC, V <sub>d</sub> , T <sub>r</sub> (5)	3.2 (-29%)	7.8 (-25%)	8.3 (-27%)	6.3 (-27%)	2.9 (-33%)	5.4 (-24%)	7.1 (-23%)	6.5 (-23%)

<sup>a</sup>WC = wet-removal coefficient; V<sub>d</sub> = dry-deposition velocity; T<sub>r</sub> = transformation rate.

ASTRAP mean predictions are also given in the table for comparison. As expected, wet deposition is most sensitive to variations in WC. Doubling the coefficient increases mean winter deposition by more than 50% and deposition in other seasons by 20% to 30%. (Lower deposition amounts are the reason that the percentage change is larger in the winter than in other seasons.) Cutting the coefficient in half reduces mean winter deposition by about 40% and deposition in other seasons by 30% to 35%. The sensitivity to doubling the  $V_d$  reduces wet deposition by 11% to 14%, and cutting the coefficient in half increases wet deposition by 9% to 16%. The model simulations of wet deposition are nearly insensitive to variations in the  $T_r$ . Doubling or halving all four model parameters simultaneously affects wet deposition to a degree somewhere between that exhibited by doubling or halving the WC and  $V_d$  individually. The increases (decreases) to wet deposition incurred by this adjustment range from 11% to 20% (7% to 10%) less than doubling (halving) WC alone.

The range in variation of ASTRAP mean predictions of wet  $\text{SO}_4^{2-}$  deposition for the parameter adjustments considered is given in Table 5.8. The minimum mean predictions across the eight seasons of simulation occur when WC is one-half its nominal value, when the diurnal/seasonal variations in  $V_d$  are twice its nominal value, and when the diurnal/seasonal variations in  $T_r$  are at one-half or at its nominal value. The maximum predictions across the eight seasons occur when the WC and variations in  $T_r$  are twice the nominal value and when  $V_d$  is one-half the nominal value. Because of model insensitivity to variations in  $T_r$ , when simulating wet deposition (due to the application of the WC to both  $\text{SO}_2$  and  $\text{SO}_4^{2-}$ ), the low WC and high  $V_d$  triple, and the high WC and low  $V_d$  triple, can be identified as the parameter sets representing the lower and upper extremes in model predictions. The strong tendency toward negative bias (overprediction) in spring and fall 1980 and fall 1981 is exhibited with mean observations in the lower fifth of the range of mean predictions. This is consistent with our previous findings made with the residual histograms, scatter plots, and time-series plots (Sec. 5.1.2).

The RDMSE or relative mean-error, expressed as a percentage of the ratio of the mean square observation and the square of the mean observation, is a fairly good indicator of the percent overall error in model predictions. Table 5.9 gives the calculated RDMSE for the nominal version of ASTRAP and the upper- and lower-extreme prediction versions just identified. The relative mean error for ASTRAP ranges from 9% (summer 1981) to 34% (fall 1980). The error in the highest PS predictor was greater than the error in the nominal ASTRAP in all cases. This error ranged from 12% (summer 1981) to 155% (winter 1981), greater than the error in the nominal version of ASTRAP. The error in the lowest PS predictor was greater than the error in the nominal ASTRAP for only five of the eight simulated seasons. This error ranged from only 15% (winter 1981) to 32% (summer 1981) greater than the error in the nominal version of ASTRAP. The error in the lowest PS predictor ranged from 10% to 20% in spring 1980 and fall 1981 to as high as 50% in winter 1980. The more favorable performance of the low WC and high  $V_d$  PS triple compared with the performance of the high WC and low  $V_d$  PS triple further supports the negative-bias (overprediction) tendency of ASTRAP, especially in spring 1980 and fall 1980 and 1981.

TABLE 5.8 Sensitivity Ranges in ASTRAP for Eight Seasons of Simulations (kg  $\text{SO}_4^{=}$ /ha)

Perform. Measure <sup>a</sup>	1980				1981			
	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
$\bar{P}_i$	2.8-7.7	5.9-14.3	6.5-16.1	4.8-12.4	2.3-7.7	4.1-9.8	5.5-13.1	5.0-11.5
WC	L - H	L - H	L - H	L - H	L - H	L - H	L - H	L - H
$V_d$	H - L	H - L	H - L	H - L	H - L	H - L	H - L	H - L
$T_r$	L/N - H	L/N - H	L - H	L/N - H	L/N - H	L - H	L - H	L/N - H
PS	8/7 - 24	8/7 - 24	8 - 24	8/7 - 24	8/7 - 24	8 - 24	8 - 24	8/7 - 24
$\bar{O}$	3.8	7.3	10.3	5.7	3.6	6.9	9.9	6.1

- <sup>a</sup> $\bar{P}_i$  = mean prediction over grid cell i  
 WC = wet-removal coefficient  
 $V_d$  = dry-deposition velocity  
 $T_r$  = transformation rate  
 PS = parameter set  
 $\bar{O}$  = mean observation over all grid cells

**TABLE 5.9 Relative Dimensionless Mean Square Error (%) and Rank Score Index in Wet Sulfate Deposition Predictions**

Performance	PS <sup>a</sup>	1980				1981			
		Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
Lowest	8	50	12	32	17	37	37	41	10
Nominal	10	22	22	13	34	22	10	9	20
Highest	24	58	61	34	84	77	23	21	53
Best overall performance		21(18)	8(3)	12(18)	14(3)	19(17)	10(11)	9(10)	7(1)
RSI <sup>b</sup> best		2.804	1.510	1.502	1.752	1.946	1.469	1.442	1.390
Worst overall performance		58(24)	61(24)	34(24)	84(24)	77(24)	37(8)	41(8)	53(24)
RSI worst		4.312	3.871	2.424	4.379	4.489	2.649	2.731	3.391

<sup>a</sup>PS = parameter set.

<sup>b</sup>RSI = rank score index.

Table 5.9 also identifies the model PS providing the combined smallest bias and scatter as measured by DMSE and IOA, and MLE and VLE,\* and the RSI for these model versions, and it gives the RDMSE. These calculations show that the nominal version of ASTRAP (or a slight modified version, with low TR) performed best for the summer and spring 1981 simulations. The model versions (high  $V_d$  and  $T_r$  and normal WC triple) performing best in summer 1980 and winter 1980 produced results that were not significantly better (only about 1%) than the nominal ASTRAP. The model versions that had performed best in spring 1980 and fall 1980 and 1981 outperformed the nominal ASTRAP by only 14%, 20%, and 13%, respectively. The RDMSE for the best and worst PS performers ranged from 7% (PS 1, fall 1981) to 21% (PS 18, winter 1980) for the best and from 34% (PS 24, summer 1980) to 84% (fall 1980) for the worst. The highest predictor, PS 24, performed the worst for six of the eight seasons. The low WC and normal  $V_d$  triple (1, 2, 3) and the normal WC and high  $V_d$  triple (16, 17, 18) performed best in spring and autumn of 1980, autumn of 1981, winter of 1980 and 1981, and summer of 1980. The normal WC and  $V_d$  triple (10, 11, 12) performed best in the two remaining seasons. The seasonal performance ranking of best, worst, and nominal PS by RDMSE will not match the same ranking by RSI, because RDMSE places a larger weight on the smallest bias error. Best and worst triples based on grouped statistics (combined seasons, paired and unpaired in time) are given in Table H.2, App. H.

The four performance measures of bias error and scatter error used to compute RSI for ASTRAP and the best and worst performance versions (Table 5.9) are given in Table 5.10, along with RSI for the nominal version of ASTRAP. These measures clearly identify ASTRAP's performance as being best for the summer seasons and for spring 1981. The measures also indicate worst performance in the winter and fall of 1980.

It is now easy to see the relative rankings among the seasons for which ASTRAP performs well and the relative rankings among the seasons for which ASTRAP performs poorly. The combined index shows the closeness of performance within the three best seasons, the two worst seasons, and the three seasons (fall 1981, winter 1981, and spring 1980) with intermediate ASTRAP performance. As indicated previously for air concentrations, additional measures are needed to determine the degree of potential improvement that can be achieved in model predictions by parameter adjustment. By decomposing mean square error into its systematic and unsystematic components, we can compute the minimum systematic MSE achievable through the factor-of-two adjustments to model parameters. (See footnote on page 84.) Table 5.10 gives the percent MSE in ASTRAP predictions that results from systematic and unsystematic causes. Also given are the computed MSE and its systematic and unsystematic parts, the minimum percent of systematic MSE achievable with PS variation, the PS producing this minimum, the MSE components for this PS, and the SERP achievable with the PS adjustments used. The largest systematic MSE occurs for model simulations of wet deposition where

---

\*As noted earlier, DMSE tends to be a bias measure and IOA tends to be a scatter measure. When ranking performance, because best performance is indicated by the largest IOA and the smallest DMSE, the inverse of IOA is added to the sum of DMSE, MLE, and VLE to get an overall ranking index, the rank score index (RSI).



**TABLE 5.10 ASTRAP Seasonal Performance and Systematic Error Reduction Potential with Parameter Variation**

Performance Measure <sup>a</sup>	Winter		Spring		Summer		Fall	
	1980	1981	1980	1981	1980	1981	1980	1981
IOA	0.493	0.673	0.629	0.822	0.850	0.849	0.565	0.700
VLE	0.262	0.242	0.121	0.116	0.138	0.100	0.215	0.135
DMSE	0.274	0.272	0.242	0.114	0.161	0.110	0.398	0.230
MLE	-0.246	-0.197	-0.351	0.015	-0.055	0.054	-0.405	-0.315
RSI	2.82	2.20	2.30	1.47	1.53	1.44	2.79	2.11
% MSE <sub>u</sub>	28.2	64.3	47.3	90.8	90.7	92.5	50.2	47.3
% MSE <sub>s</sub>	71.8	35.7	52.7	9.2	9.3	7.5	49.8	52.7
RMSE	2.17	2.00	4.28	2.38	4.32	3.07	4.39	3.38
RMSE <sub>u</sub>	1.16	1.61	2.94	2.27	4.11	2.96	3.11	2.32
RMSE <sub>s</sub>	1.84	1.20	3.11	0.72	1.32	0.84	3.10	2.45
Minimum % MSE <sub>s</sub>	71.7	33.2	31.1	7.1	6.2	2.4	35.6	12.1
PS	11	18	17	26	18	26	5	3
RMSE	2.16	1.76	3.20	2.86	3.91	3.69	2.47	1.64
RMSE <sub>u</sub>	1.15	1.44	2.66	2.75	3.78	3.65	1.98	1.54
RMSE <sub>s</sub>	1.83	1.01	1.79	0.76	0.97	0.57	1.48	0.57
SERP <sup>b</sup>	<1%	<3%	22%	<3%	3%	5%	14%	41%

<sup>a</sup>IOA = index of agreement  
VLE = variance logarithmic error  
DMSE = dimensionless mean square error  
MLE = mean logarithmic error  
RSI = rank score index  
MSE<sub>u</sub> = mean square error unsystematic  
MSE<sub>s</sub> = mean square error systematic  
RMSE = root mean square error  
RMSE<sub>u</sub> = root mean square error unsystematic  
RMSE<sub>s</sub> = root mean square error systematic  
PS = parameter set

<sup>b</sup>Systematic error reduction potential is approximate.

ASTRAP performance is fair to poor. Almost 72% of the winter 1980 simulation error is systematic. Although this represents the largest percentage of systematic error of the eight seasonal wet-deposition simulations, the SERP is less than 1%. Because the wet chemistry monitoring network (Acid Deposition Monitoring system) was just really getting started in late 1979 and early 1980, systematic error in sample collection and analysis could be a significant cause for poor model performance over this period. The SERP is also small for the winter, spring, and summer of 1981. The difference in these simulations is that the apparent systematic error in nominal ASTRAP predictions and is less than 10% of the total apparent error. In these cases, the observational error may not, with the exception of the winter 1981, be as significant a contributor to the systematic error as it was in the winter 1980 simulations. The basis for this exception is the documented systematic undercatch error in precipitation sampling, which is a result of wind field deformation above the rain gauge orifice (more predominant with the elevation of the orifice above the ground), wetting losses on internal sampler walls, evaporation losses, snow blowing and drifting, and splash-out or splash-in. The systematic undercatch in precipitation has been estimated, from experimental data collected in Europe and the USSR with the pit gauge as the reference, to vary from 3% to 30% annually and as much as 50% or more for individual episodes (Rodda et al. 1985 and 1986, Sevruk 1982). The systematic undercatch, although more significant in the winter, may also contribute to a good portion of the systematic error in spring 1980 and fall 1980 and 1981. SERP ranged from 14% to 41% for these periods, which may be caused not only by the occasionally significant snowfall in Canada and northern states in these periods but also by the systematic undercatch of rain resulting from the same physics causing snow undercatch in nonpit rain gauges. The very small systematic error and SERP for spring 1981 and summer 1980 and 1981 may indicate that not much model performance improvement can be achieved over these periods except through model reformation or the preparation of a high-resolution (spatial and temporal) meteorology, emissions, and wet-deposition sampling data base.

### 5.3 ERROR DECOMPOSITION AND SPATIAL ERROR PATTERN ANALYSIS

The preceding discussion relied primarily on the use of distributional statistics to quantify and express apparent error in model predictions. Error was expressed in terms of residuals, scatter or variance, and mean squares, including the systematic and unsystematic portions of the mean square error (MSE). Although we now have a better picture of how ASTRAP performs, we are still missing some key elements needed to understand this performance. We need to be able to express the apparent error in terms of its bias, temporal, and spatial error components, and to graphically display and quantify the spatial error component. We propose to do this by the analysis of variance through the decomposition of MSE and through analysis of variance with regression analysis and decomposition of explained variance (in Sec. 5.3.1), and through display of spatial patterns that are optimized in an MSE sense (method described in Sec. 5.3.2).

#### 5.3.1 Separation and Computation of Bias, Temporal, and Spatial Error Components

The MSE for a set of space-time observations and predictions can be decomposed into three parts through analysis of variance (ANOVA) (Ball 1986). This decomposition of

error could be important in understanding what and where the weak links in the model and the model input data may be.

Suppose we have a series of predictions  $P_{ik}$  and observations  $O_{ik}$  at  $M$  locations (index  $i$ ). At each location  $i$  (receptor grid,  $130 \times 130$  km), there are  $K_i$  observations in time (index  $k$ ), for a total of  $N = \sum K_i$  observation/prediction pairs. Equation 4.5 for MSE in Section 4.2.1 can now be written as:

$$MSE = \frac{1}{N} \sum_{i=1}^M \sum_{k=1}^{K_i} [O_{ik} - P_{ik}]^2 \quad (5.1)$$

The mean values over time at each location  $i$  are given by:

$$\langle O_i \rangle = \frac{1}{K_i} \sum_{k=1}^{K_i} O_{ik} \quad (5.2)$$

$$\langle P_i \rangle = \frac{1}{K_i} \sum_{k=1}^{K_i} P_{ik} \quad (5.3)$$

The overall mean values (over time and space) are given by:

$$\langle\langle O \rangle\rangle = \frac{1}{N} \sum_{i=1}^M \sum_{k=1}^{K_i} O_{ik} \equiv \frac{1}{N} \sum_{i=1}^M K_i \langle O_i \rangle \quad (5.4)$$

$$\langle\langle P \rangle\rangle = \frac{1}{N} \sum_{i=1}^M \sum_{k=1}^{K_i} P_{ik} \equiv \frac{1}{N} \sum_{i=1}^M K_i \langle P_i \rangle \quad (5.5)$$

Equations 5.2 through 5.5 can be used to define, for ease of notation, the mean residual at location  $i$  over all time, and the overall mean residual or bias over the entire field of values as:

$$\langle r_i \rangle = \langle O_i \rangle - \langle P_i \rangle \quad (5.6)$$

$$\langle\langle r \rangle\rangle = \langle\langle O \rangle\rangle - \langle\langle P \rangle\rangle \quad (5.7)$$

In terms of these definitions, we can rewrite Eq. 5.1 as the sum of three terms:

$$MSE = \frac{N-M}{N} MSTE + \frac{M-1}{N} MSSE + \langle\langle r \rangle\rangle^2 \quad (5.8)$$

where MSTE is the mean square temporal error given by:

$$MSTE = \frac{1}{N-M} \sum_{i=1}^M \sum_{k=1}^{K_i} [O_{ik} - P_{ik} - \langle r_i \rangle] \quad (5.9)$$

which can be related to the ANOVA notation as the sum of squares within groups, as  $SS_g = (N - M) MSTE$ .

MSSE is the mean-square spatial error given by:

$$MSSE = \frac{1}{M-1} \sum_{i=1}^M K_i [\langle r_i \rangle - \langle\langle r \rangle\rangle]^2 \quad (5.10)$$

which can be related to ANOVA notation as the sum of squares between groups,  $SS_b = (M - 1) MSSE$ .

In order to make the comparison of error totals across seasons meaningful, it is necessary to normalize the MSE in Eq. 5.8 by the product of the interannual mean observation and prediction (the combined mean value of like seasons of different years). Equation 5.8 then becomes:

$$DMSE = \frac{N - M}{N \cdot \bar{O} \cdot \bar{P}} MSTE + \frac{M - 1}{N \cdot \bar{O} \cdot \bar{P}} MSSE + \frac{1}{\bar{O} \cdot \bar{P}} \langle\langle r \rangle\rangle^2 \quad (5.8a)$$

We now have a means to separate error into its temporal, spatial, and bias components. These components provide a means to represent the composition of error in model simulations for particular seasons. The temporal component is a measure of the ability of the model and model input data to account for interseasonal variations in meteorology and emissions. The significance of this measure may be highly dependent on the number of available data years and can be greatly influenced by meteorological variability from year to year. The spatial component is a measure of the ability of the model and data base to accurately represent the spatial patterns in deposition and air concentrations. These patterns are highly contingent on the model representation of wind and precipitation fields. The significance of this measure is dependent upon the spatial distribution and the number of receptor sites. Finally, the bias component is an expression of the residual error or the overall systematic bias over the entire field of values.

Dividing each of the terms in Eq. 5.8a by DMSE provides a way to express the temporal, spatial, and bias error components as a percentage of the total error. These

components (TE, SE, and BE) are provided in Table 5.11, along with the bias, temporal, and spatial error ratios (BER, TER, SER), and the ratio of spatial to temporal error (STER). The first three ratios are error fractions of the interseasonal mean (for BER) and the interseasonal mean square observation. The STER values given provide some idea of the relative importance of spatial error to temporal error. An F-test of this ratio can be interpreted as a test of the hypothesis that the spatial differences are not random but are statistically significant. One can expect real or significant differences in the magnitude of bias errors among sites or regions when STER is larger than about 1.5. This suggests that the variations in error among regions for the winter, summer, and fall simulations are significant. These ratios can give a misleading picture, however, if the bias error component is a significant fraction of the total error (i.e., greater than 20%). The error ratios show that the spatial and temporal error ratios are 2 to 18 times larger than the BER. An examination of the contributions to total error shows that the spatial error dominates, accounting for over 70% of the total in the winter, spring, and summer. These statistics give the first indication of the apparent dominance of the spatial error component and suggest that the model is not doing well in representing spatial patterns (see the following discussion on explained variance and the discussion in Section 5.3.2). The autumn simulations show dominance of the spatial-error and bias-error components, with a small temporal-error component. The relatively small temporal-error component across all seasons, particularly winter and autumn, may be the result of the statistically small number of seasons considered in the analysis. With only two years of data available for analysis, only two points contribute to the variance at each site. Because of the relatively large size of the total error in autumn, the autumn spatial error ( $DMSE_s = 0.148$ ) is only slightly larger than the spatial error for the spring ( $DMSE_s = 0.122$ ) and summer ( $DMSE_s = 0.110$ ). In addition to representing the largest percentage of the total error, the spatial error for the winter simulations ( $DMSE_s = 0.216$ ) is 60% to 95% larger than the spatial error for the other seasons. The absolute bias error for the autumn simulations is from 4 to more than 100 times larger than the bias error in the spring, summer, and winter simulations.

We previously computed the explained variance of model predictions over individual periods or seasons (Section 5.1.2). Now we will report the explained variance in terms of a temporal and spatial component. To do this, it is useful to pick up aspects of ANOVA and combine them with a regression analysis (Ball 1987). The regression model definitions for the total sum of squares of the observations (SSTO), the sum of squares error (SSE), and the sum of squares regression (SSR) follow:

$$SSTO = \sum_{i,k} (O_{ik} - \langle\langle O \rangle\rangle)^2 = (N - 1) \sigma_o^2 \quad (5.11)$$

$$SSE = \sum_{i,k} (O_{ik} - \hat{P}_{ik})^2 = N \cdot (MSE) \quad (5.12)$$

$$SSR = \sum_{i,k} (\hat{P}_{ik} - \langle\langle O \rangle\rangle)^2 = (N - 1) \cdot \sigma_{\hat{P}}^2 \quad (5.13)$$

**TABLE 5.11 Temporal, Bias, and Spatial Error in ASTRAP Predictions of 1980 and 1981 Wet Sulfate Deposition**

Season	M <sup>d</sup>	N <sup>e</sup>	$\bar{O}/\bar{P}^f$ (kg/ha)	DMSE <sup>g</sup>	Bias Components <sup>a</sup>			Temporal Components <sup>b</sup>			Spatial Components <sup>c</sup>			
					MBE (kg/ha)	BER	BE (%)	RMSTE (kg/ha)	TER	TE (%)	RMSSE (kg/ha)	SER	SE (%)	STER
Winter	46	57	3.62/ 4.38	0.273	-0.76	0.21	13.3	1.29	1.28	7.4	2.08	1.09	79.3	2.64
Spring	59	77	7.12/ 8.29	0.174	-1.18	0.17	13.5	2.62	2.82	17.0	3.13	1.05	70.3	1.29
Summer	70	93	9.69/ 10.03	0.136	-0.34	0.04	1.0	2.88	0.74	11.6	3.87	0.58	80.8	2.33
Autumn	73	101	5.79/ 8.47	0.302	-2.68	0.46	48.4	1.73	0.70	2.8	3.18	1.43	49.1	6.93

<sup>a</sup>Bias Components

- Mean Bias Error (MBE)
- Bias Error Ratio:  $BER = MBE/\bar{O}$
- Bias Error:  $BE = (MBE)^2/DMSE * \bar{O} * \bar{P}$

<sup>b</sup>Temporal Components

- Root Mean Square Temporal Error (RMSTE)
- Temporal Error Ratio:  $TER = MSTE/MSTO$
- Temporal Error:  
 $TE = (N - M)/N * MSTE/DMSE * \bar{O} * \bar{P}$

<sup>c</sup>Spatial Components

- Root Mean Square Spatial Error (RMSSE)
- Spatial Error Ratio:  $SER = MSSE/MSO$
- Spatial Error:  $SE = (M - 1)/N * MSSE/DMSE * \bar{O} * \bar{P}$
- Spatial-Temporal Error Ratio:  $STER = MSSE/MSTE$

<sup>d</sup>M = number of sites producing at least one observation/prediction pair.

<sup>e</sup>N = number of observation/prediction pairs.

<sup>f</sup> $\bar{O}/\bar{P}$  = mean observation/mean prediction.

<sup>g</sup>DMSE = dimensionless mean square error.

where:

$O_{ik}$  = observational field,

$\hat{P}_{ik}$  = regression of model predictions, and

$\hat{P}_{ik} = a + b O_{ik}$ , the linear least-squares regression model.

We now can define the coefficient of determination ( $R^2$ ) as:

$$R^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO} = \frac{\frac{\sigma_{\hat{P}}^2}{2}}{\frac{\sigma_o^2}{2}} \quad (5.14)$$

$$= 1 - \frac{SSE}{SSTO} \quad (5.15)$$

Using Eqs. 5.14, 5.12, and 5.13,  $R^2$  can be defined in terms of MSE:

$$R^2 = 1 - \left( \frac{N}{N-1} \right) \frac{MSE}{\sigma_o^2} \quad (5.16)$$

The sigma ratio in Eq. 14 is valid for a regression model but not valid when  $\hat{P}_{ik}$  is not a least-squares regression fit to  $O_{ik}$ . In other words, it is no longer true that  $SSE + SSR = SSTO$ . This is because the observational data set,  $O_{ik}$ , does have measurement error. (It should also be noted that the overall mean of observations,  $\langle\langle O \rangle\rangle$ , is also the mean of the regression model,  $\langle\langle \hat{P} \rangle\rangle$ , but this is not true when there is an overall bias.) Because of this error, to not penalize the model, a slightly modified definition of  $R^2$  seems appropriate. First, a new sum of square error is defined as follows:

$$SSE' = \sum_{i=1} [O_{ik} - \langle\langle O \rangle\rangle - (\hat{P}_{ik} - \langle\langle \hat{P} \rangle\rangle)]^2 \quad (5.17)$$

Removing bias from Eq. 5.12 yields:

$$SSE' = N [MSE - (BIAS)^2] \quad (5.18)$$

Using Eqs. 5.17 and 5.18, the bias term can be defined as:

$$BIAS = \langle\langle O \rangle\rangle - \langle\langle \hat{P} \rangle\rangle$$



Now the explained variance in model predictions can be defined in terms of the total variance in the observations minus the unexplained variance or error in observations as:

$$R'^2 = \frac{SSTO - SSE'}{SSTO} = 1 - \frac{SSE'}{SSTO}$$

Using Eqs. 5.18 and 5.14, the explained bias-corrected variance (EBCV) can be expressed as:

$$R'^2 = 1 - \left( \frac{N}{N-1} \right) \frac{MSE - (BIAS)^2}{\sigma_o^2} \quad (5.19)$$

By using the following definitions for mean square temporal observation (MSTO), mean square spatial observation (MSSO), and observed variance ( $\sigma_o^2$ ), and Eqs. 5.9 and 5.10, the explained variance can now be decomposed to a solely spatial, plus a solely temporal component (the bias component has been removed).

$$MSTO = \frac{\sum_{i=1}^N [O_i - \langle O_i \rangle]^2}{N - M} \quad (5.20)$$

$$MSSO = \frac{\sum_{i=1}^N K_i [\langle O_i \rangle - \langle \langle O \rangle \rangle]^2}{M - 1} \quad (5.21)$$

$$\sigma_o^2 = \frac{N - M}{N - 1} MSTO + \frac{M - 1}{N - 1} MSSO \quad (5.22)$$

With some manipulation, it can be shown that the EBCV ( $R'^2$ ) can be expressed as the sum of a temporal component and a spatial component:

$$EBCV = \left( \frac{N - M}{N - 1} \right) \cdot \frac{MSTO}{\sigma_o^2} R_t^2 + \left( \frac{M - 1}{N - 1} \right) \cdot \frac{MSSO}{\sigma_o^2} R_s^2 \quad (5.23)$$

where:

$$R_t^2 = \frac{MSTO - MSSE}{MSTO}, \text{ the temporal error component of EBCV, and}$$

$$R_s^2 = \frac{MSSO - MSSE}{MSSO}, \text{ the spatial error component of EBCV.}$$

As these components approach 1.0 (as MSTE and MSSE approach zero), EBCV approaches 1.0, and the model can be said to explain all the variance in the observations.

Now the EBCV and its temporal and spatial components can be computed for interseasonal comparisons of wet sulfate deposition observations and predictions in 1980 and 1981. The results from these computations are summarized in Table 5.12. The table provides the EBCV and the size of its temporal (TVE) and spatial (SVE) components, all expressed as percentages of the total explained variance in observations. The temporal and spatial error components ( $R_t^2$  and  $R_s^2$ ) are also given, ranging from less than or equal to zero (mean square error is equal to or greater than the mean square observation) to 1.0 (mean square error equals zero). As the spatial or temporal error components get larger, the spatial or temporal error variances get smaller. The data variances show that the model's ability to explain variance in summer simulations (over 40%) is substantially better than its ability to explain variance in winter, spring, and autumn simulations. This difference in ability results from the model's improved ability to explain spatial variance in the summer simulations examined. More than 90% of the EBCV in the summer results from the ability to explain spatial patterns. The negative EBCV values indicate that the model does not do very well in explaining the observed interannual variance in nonsummer seasons. The computed SVE ranged from -43% for autumn simulations to

**TABLE 5.12 Explained Variance in ASTRAP Predictions**

Season	TVE <sup>a</sup> (%)	$R_t^2$	SVE <sup>b</sup> (%)	$R_s^2$	Total EBCV <sup>c</sup> (%)
Winter	-2.0	-0.28	-8.4	-0.09	-10.4
Spring	-14.9	-1.83	-4.1	-0.04	-19.0
Summer	2.6	0.26	38.3	0.43	40.9
Autumn	-0.9	-0.09	-43.2	-0.48	-44.1

<sup>a</sup>TVE = Temporal variance explained, expressed as a percent of the total explained variance:

$$TVE = \left( \frac{N - M}{N - 1} \right) \cdot \frac{MSTO}{\sigma_o^2} \cdot R_t^2 \cdot 100\%$$

<sup>b</sup>SVE = Spatial variance explained, expressed as a percent of the total explained variance:

$$SVE = \left( \frac{M - 1}{N - 1} \right) \cdot \frac{MSSO}{\sigma_o^2} \cdot R_s^2 \cdot 100\%$$

<sup>c</sup>EBCV = Explained bias-corrected variance.

slightly over 38% for summer simulations. The model appears to not do well in explaining temporal variations, but this result must be viewed with caution because of the limited amount of data available (only two years) at the initiation of our study.

### 5.3.2 Spatial Air Concentration and Deposition Error Patterns

ASTRAP shows a limited ability to simulate spatial patterns of observed wet sulfate deposition. To see this more clearly, a technique to display spatial patterns in observations and predictions is needed. Several contour mapping techniques are available for this purpose, such as distance weighting, ordinary and generalized linear least squares, and polynomial least squares. Most classical techniques for computing surface contours to determine spatial trends are based on assumptions that the observed data can be represented by a polynomial or piecewise polynomial surface through a least-squares fit of the data and that the deviations from a smooth surface are assumed to be random errors (Goodin et al. 1979, Ripley 1981, McLarin 1974). Distance weighting techniques are based on the assumptions that the physical and chemical processes that affect a specific site also affect points nearby and that the effect of the specific site on nearby points can be assumed to be a function of the distance of separation (e.g., inverse distance, inverse distance squared). Ordinary least-squares techniques approximate drift between data points with a linear or polynomial least-squares regression. These techniques suffer from a number of weaknesses such as subjectively determining the weight function, neglecting the possibility that the interpolation errors may have covariance, and not providing an estimate of interpolation errors.

The contour mapping technique chosen should have several key attributes. At a minimum, it should preserve the main characteristics of the data by smoothing through inherent random data variability. The technique should retain the important spatial features of the data and should not be unduly influenced by values at single points. Ideally, the degree of variability in the data should influence the degree of smoothing, and the uncertainty in the interpolation estimates that are produced should be provided (Clark et al. 1987a). These desired attributes are for the most part inherent to a geostatistical interpolation technique known as kriging. Kriging is designed to minimize the overall variance between the true values and the estimated or interpolated values. The contours produced by kriging are, under certain conditions, designed to be optimal in a mean square sense. The weighting functions used in the interpolation are a function of the spatial distribution of data points and the inherent variability in the data. The approach is based upon the theory of regionalized variables, which is designed to mathematically describe geophysical properties distributed in space and/or time and provide an appropriate means for solving spatial estimation problems (Matheron 1971). The technique was used extensively in the 1960s and 1970s in geology (mineral exploration), oceanography, meteorology (rainfall and geopotentials), hydrology (water table heights), radiochemistry (geographical distribution of radionuclides), and more recently in the analysis of spatial trends in acid-precipitation data (Finkelstein 1983, Bilonick 1985, Eynon and Switzer 1983).

Kriging is not a single method but a family of interpolation techniques. Two basic approaches, universal and simple, are most often applied. The simple or ordinary kriging method has been used almost exclusively to spatially extrapolate acid-precipitation data. This method is dependent on the assumption that the underlying process is stationary (i.e., that the expected value of the process does not vary appreciably over distances between the data points and the interpolation points). The drift or trend between data and interpolation points is assumed to be constant. If the trend cannot be assumed constant (i.e., is nonstationary), a linear combination of predictors (e.g., polynomial terms) is introduced into the system of equations used to determine interpolation weights. This introduction of terms to account for a varying trend is known as the universal kriging method. Since the method relies on the preselection of the functional form of the trend and the error covariance, if the wrong form is chosen the bias in the interpolated variance may be worse than that obtained from simple kriging (Seilkop 1983). Oden (1986) has developed a graphical inspection and a spatial autocorrelation method to test the adequacy of the trend and/or error covariance functional form and to provide a means to alter the functions, if necessary, through parameter variation. This approach is sometimes referred to as generalized covariance kriging (Dennis and Seilkop 1986). A detailed description and the full mathematical treatment of the spatial autocorrelation fitting of the drift estimator used in the covariance kriging algorithm can be found in Oden (1986). The computational details of simple and universal kriging are provided in App. K.

Both simple and generalized covariance kriging (Oden 1986) were investigated for evaluating spatial patterns of observed and predicted  $\text{SO}_2/\text{SO}_4$  air concentrations, wet sulfate deposition, and the difference in those patterns. Due to difficulty in properly interpreting anomalies appearing in the generalized covariance kriging output, the results are reported in App. I without discussion. Our analysis is based upon the comparison of spatial patterns of simple kriging results. The level of spatial analysis used in this study is not as quantitative as the analysis used in the ISDME study (Clark et al. 1987a). A scaled down analysis was done for three reasons. First, because of the lack of grid point ASTRAP predictions, spatially kriged predictions (from observed evaluation sites to these grid points) could not be compared with the true ASTRAP grid point values. Without such a comparison, we could not fully determine how well the kriging routine reproduced the true model predicted spatial patterns. The counterpoint would be that if the predicted spatial patterns were determined on this basis, the kriged observed spatial patterns would not have the same advantage of a spatially uniform and dense set of interpolation data points. Second, the sophisticated statistical mathematical technique used to derive the kriged weighting functions can often overlook the fact that there is little physical justification for a model represented by  $Z(X) = \sum w_i Z_i(X)$ , where  $w_i$  is statistically derived spatial weighting functions\* (Venkatram and Pleim 1985). A counter argument can be made, stating that the physical processes that govern wet-deposition spatial patterns (i.e., precipitation and wind fluctuation) are highly stochastic and that a

---

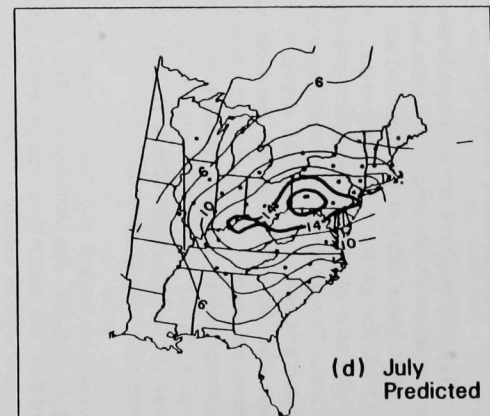
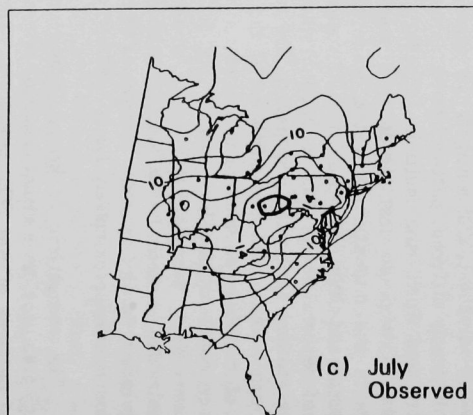
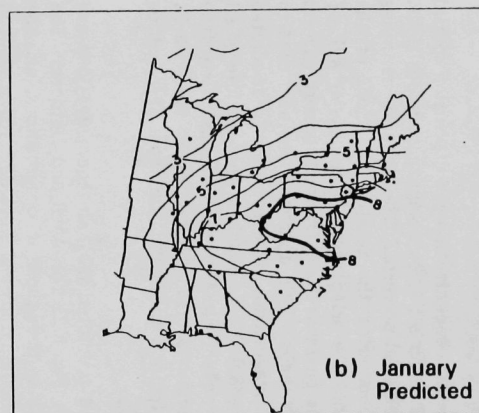
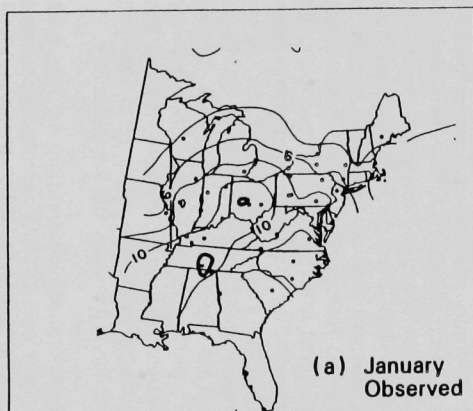
\*The spatial weighting function provides interpolation weights that depend on the covariance structure of the regionalized variable,  $Z(X)$ . The function  $Z_i(X)$  is a random function composed of a stochastic component and a deterministic or trend component.

model that attempts to account for this is therefore justified. Finally, sufficient resources were not available to conduct an analysis at the level used in the ISDME study. Because ASTRAP was one of the models evaluated in this study, however, more quantitative results could be used to support our data interpolation and analysis.

The ability of the model to reproduce the air concentration and deposition spatial patterns inherent in the observations is assessed in our study by comparing the positions, shapes, and magnitudes of the observed and predicted patterns. The differences in the position of the patterns are determined by comparing the relative locations of the observed and predicted high-deposition fields and the orientation of the major axis of each field. The differences in the shapes of the patterns is determined through careful inspection and comparison of the kriged contour maps. This approach was also used to compare the differences in the locations and magnitudes of observed and predicted maximum air concentrations and deposition. It should be emphasized that our use of all these comparative measures is qualitative but nevertheless useful when used with the more quantitative results presented in the previous section of this report (Sec. 5.3.1). Refer to the ISDME report (Clark et al. 1987a) for how these and other spatial measures can be quantitatively treated.

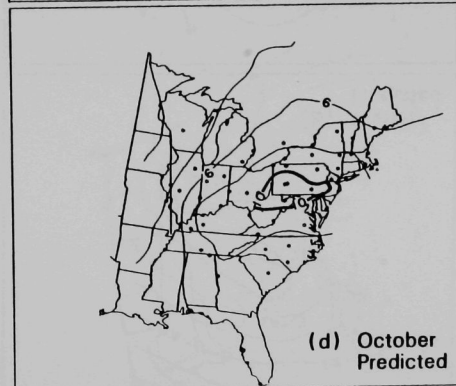
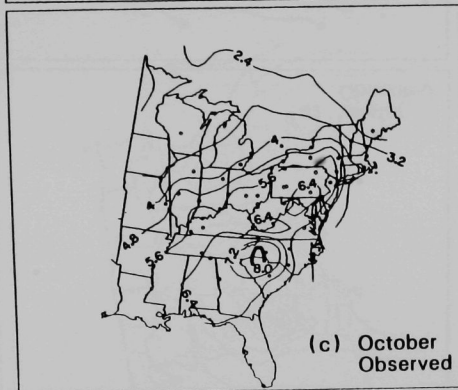
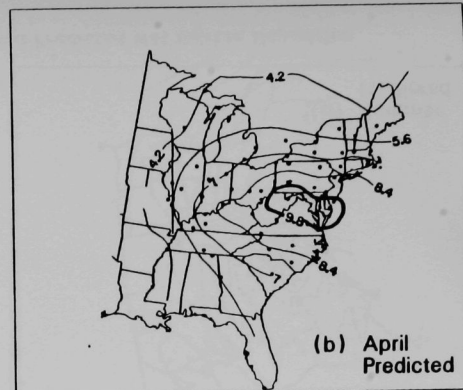
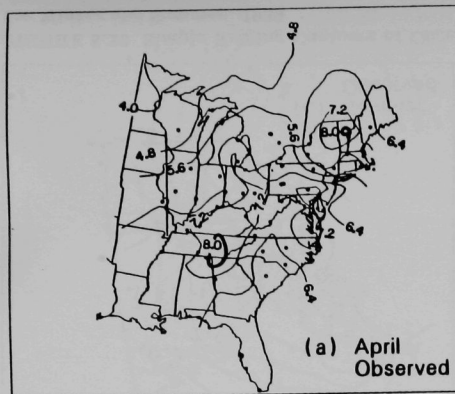
Simple kriged contours of observed and predicted  $\text{SO}_4^-$  air concentrations are shown in Fig. 5.20 for January and July and in Fig. 5.21 for April and October. The contours of the predictions are generally smoother than those of observations, and the predictions tend to exhibit a single "hot spot" (maximum area). More detail is present in the observed contours, with a tendency toward multiple hot spots. The orientation of the major axis of the prediction contours is east-west for all months but July, when it is southwest-northeast. In contrast, the observed contours are shaped more irregularly, with distinctly different patterns for each month. Multiple observed hot spots occur in January and April, and multiple secondary hot spots occur in October. Major axis orientation is predominantly southwest-northeast. The January predicted maximum occurs in the mid-Atlantic states, while the observed January maximum occurs in Tennessee and Ohio. The predicted July maximum is reasonably close to but somewhat east of (approximately 300 to 350 km) the observed maximum. The two observed (Tennessee and Vermont) April maxima are represented by a single predicted maximum area, approximately 450 km northeast of the observed maxima in Tennessee and 450 km south of the observed maxima in Vermont. A rather broad area of predicted maximum October concentration is shown in Fig. 5.21(d) (southern Pennsylvania, New Jersey, northern West Virginia, Maryland), while a more localized area of maximum concentration is shown in Fig. 5.21(c) for observations (southwestern North Carolina and northwestern South Carolina). The ratios of the magnitude of the predicted to observed maximums are approximately 1.3 in April and October, 1.0 in July, and 0.7 in January.

The comparison of kriged (simple) predictions and observations in Figs. 5.22 through 5.25 provides a visual means of assessing how well ASTRAP reproduces observed spatial patterns of wet  $\text{SO}_4^-$  deposition. The general spatial features of the observed patterns are more heterogeneously detailed than are the predicted patterns, showing multiple peak deposition areas in the summer 1981 and varied contour orientations and shapes. In contrast, the predicted patterns show less variation in characteristics from



**FIGURE 5.20 Simple Kriging Contours of Observed and Predicted Sulfate Air Concentrations for January and July, 1978**

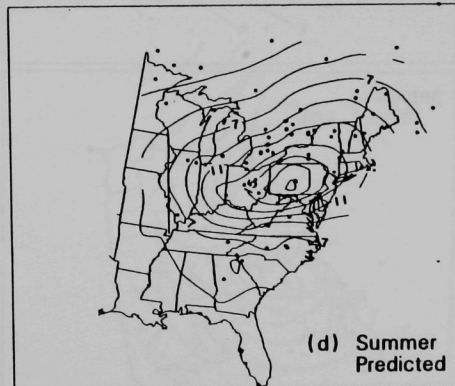
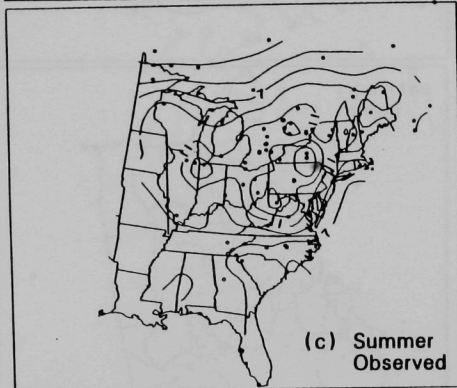
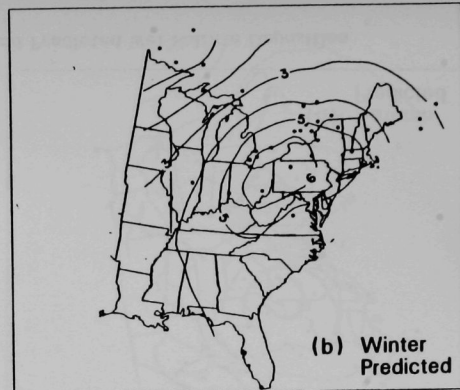
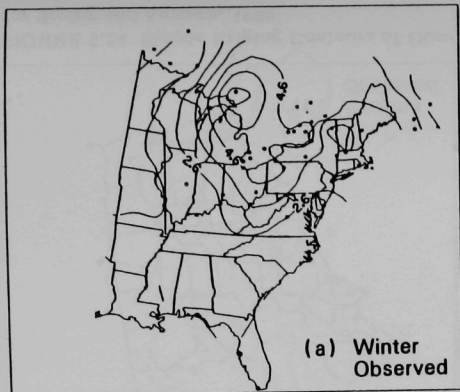




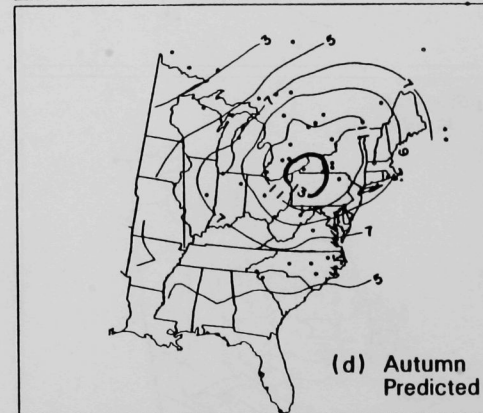
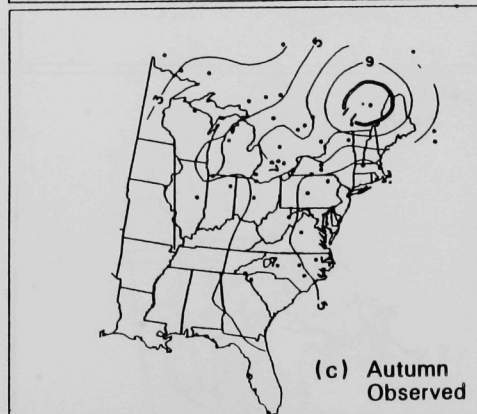
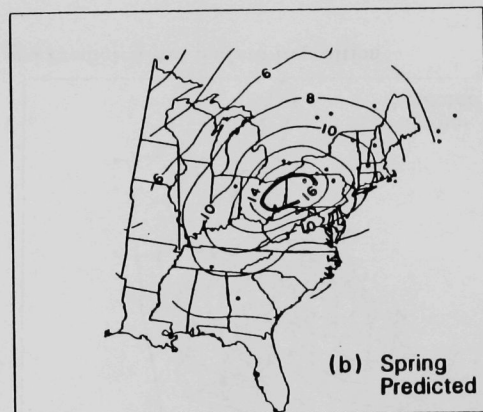
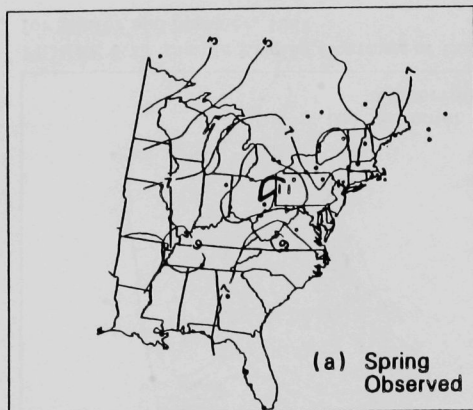
**FIGURE 5.21 Simple Kriging Contours of Observed and Predicted Sulfate Air Concentrations for April and October, 1978**



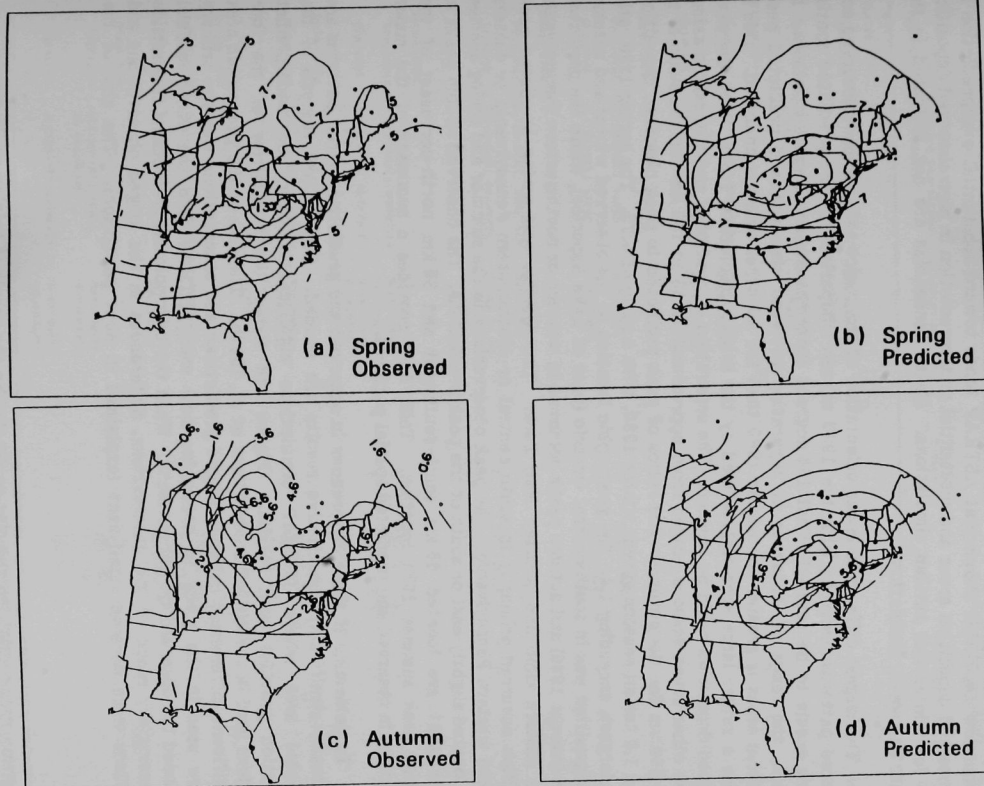




**FIGURE 5.23 Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Winter and Summer, 1981**



**FIGURE 5.24 Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Spring and Autumn, 1980**



**FIGURE 5.25 Simple Kriging Contours of Observed and Predicted Wet Sulfate Deposition for Spring and Autumn, 1981**

season to season. The irregular contours for 1980 winter predictions and observations probably result, in part, from the sparse spatial density of observation sites available (only 25 data points to represent eastern North America).<sup>\*</sup> A comparison of patterns of high and low deposition shows that ASTRAP tends toward substantial overprediction in low observed deposition areas and substantial underprediction in high observed deposition areas for winter and autumn simulations. These tendencies are not indicated in the summer and spring simulations.

The largest difference in the orientation of the major axis between observed and predicted patterns occurred with the 1980 winter comparisons. The predicted spatial pattern seems to be rotated a full 180 degrees from the observed spatial pattern. In addition, the range of distances of separation between observed and predicted peak deposition areas is greater for winter 1980 than for all other seasons analyzed. This is because a rather large area was defined by the kriged prediction, from which the peak predicted deposition could be located. The separation of predicted and observed maxima for the other seven seasons ranged from approximately 50 to 100 km for the summer to about 800 km for the autumn. The ratios of peak predicted to peak observed deposition exceed 1.0 for all seasons except winter 1980, with a ratio of 0.75. The spring 1980 ratio is the largest, exceeding 1.4. The geographic location of the observed winter and autumn peak deposition was in south central Ontario (east of Lake Superior), while the predicted winter (except 1980) and autumn peaks occurred in western or northwestern Pennsylvania and/or eastern Ohio and western New York. The summer and spring predicted peak deposition occurred primarily in south central or southwestern Pennsylvania or eastern Ohio and western Pennsylvania. The peak observations in the summer and spring seemed to be located slightly east or south of the peak predictions. The observed double peaks in summer 1981 are located 50 km south-southwest and 50 km north-northeast of the predicted peak summer 1981 location. Table 5.13 provides a summary of the major differences in observed and predicted spatial patterns.

To determine if these differences in observed and predicted spatial patterns are statistically significant, quantitative results are needed. The ISDME analysis of the geographic area where ASTRAP simulations significantly over- or underpredict observation provides such results. Figure 5.26 shows the boundaries of the four subregions used in the ISDME study (Clark et al. 1987b). The size of the geographic area of significant differences between kriged predictions and observations was the primary measure used in the study to determine how well ASTRAP (and ten other models) reproduced the observed spatial patterns. When the interpolated predictions fell outside the uncertainty range of the observations, differences in the kriged observations and predictions were deemed significant (explained in next paragraph). The size of the

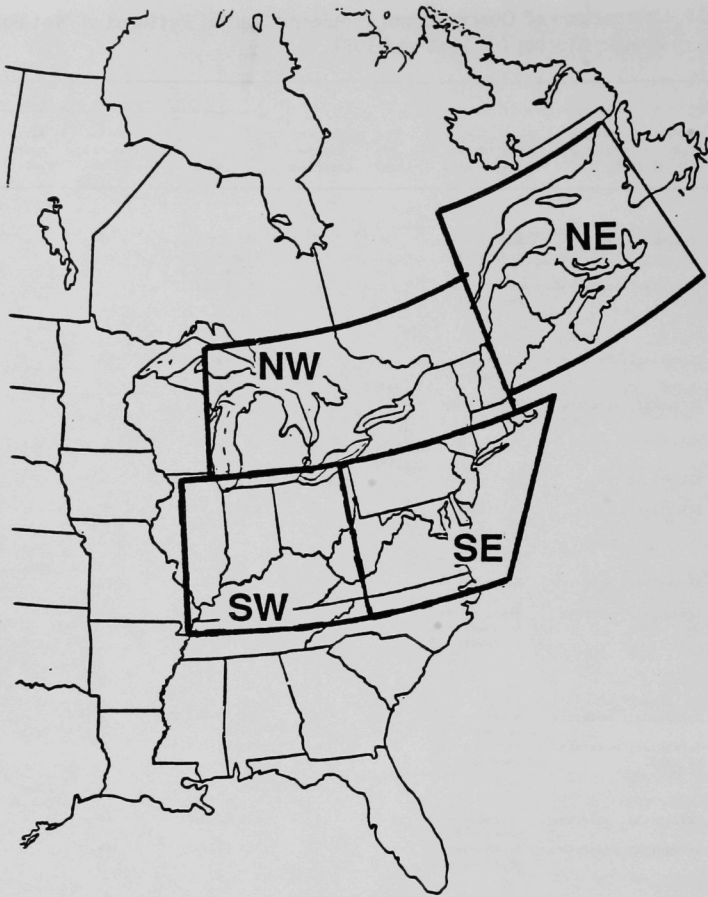
---

<sup>\*</sup>In addition, this spatial irregularity could also be the result of kriging predictions based on validation sites rather than much more uniform and densely distributed model grid point coordinates. The influence of this on kriged patterns can be seen in the more detailed (winter) and the smoother and more regularly shaped concentric ellipses produced for the 1980 seasonal ASTRAP predictions in the ISDME study (Clark et al. 1987a).

**TABLE 5.13 Comparison of Observed and Predicted Spatial Patterns of Wet Sulfate Deposition — Simple Kriging for 1980 and 1981**

Season	Pattern Shape; Relative Deposition Gradient	Orientation of Major Axis	$\Delta^a$ (deg)	No. of Maxima Locations	Location and Magnitude of Maxima			
					Area	Est. Mag. (kg SO <sub>4</sub> /ha)	Est. Separ. (km)	Geography
Winter 1980			180° (8.7°)				300- 1600	
Predicted	Irregular; small	Northeast- southwest		1	REG III; VI.b and d; VII.a and b	6		Indeterminant
Observed	Irregular; localized large	Northwest- southeast		1	REG VI.c.4.7			S.E. Ontario
Winter 1981			90°				450- 950	
Predicted	Elliptical; small	Northeast- southwest		1	REG VI.a, b and c	8		E. Ohio- W. N.Y.- W. Penn.
Observed	Irregular; moderate	North-south		1	REG I.d.7.	7		Ontario-E. Lake Superior
Summer 1980			15° (8.5°)				10-350	
Predicted	Elliptical; large	East-west		1	REG VI.b.8; d.2	24		S.W. Penn.
Observed	Elliptical; large	East-west		1	REG IV.b.2, 3,5,6,8	20		S.E. Ohio
Summer 1981			45°				240- 360	
Predicted	Elliptical; moderate	East-west		1	REG VI.d.2,5	20		S. Cen. Penn.
Observed	Irregular; moderate	Seems to be northeast- southwest		2	REG VI.c.6,9; b.9; d.3; d.4.7	16		S. Cen. N.Y.- N. Cen. Penn.- N.W. Va.
Spring 1980			20° (3.1°)				10-350	
Predicted	Elliptical; moderate	Northeast- southwest		1	REG VI.b.7,8; d.1,2,4	17		E. Ohio- E. Penn.
Observed	Irregular; moderate to small	Northeast- southeast		1	REG VI.b.4,5	12		E. Ohio- W. Penn.
Spring 1981			15°				120- 240	
Predicted	Elliptical; moderate	Northeast- southeast		1	REG VI.b.8; d.2	14		S.W. Penn.
Observed	Irregular; moderate	Northeast- southeast		1	REG VI.d.3	14		N.W. Va.
Autumn 1980			15° (16.6°)				600- 1200	
Predicted	Elliptical; moderate	Northeast- southeast		1	REG VI.c.3,6; d.1,2,5,6	14		W. N.Y.- N.W. Penn.
Observed	Irregular; moderate	Northeast- southeast		1	REG VIII.b.7,8; IX.c.1.2	12		S.E. Ontario- S. Quebec
Autumn 1981			90°				720- 960	
Predicted	Elliptical; small to moderate	Northeast- southeast		1	REG VI.d.1,2, 4,5; b.4, 5,7,8	7		E. Ohio- W. Penn.
Observed	Irregular; moderate	Northeast- southeast		1	REG V.b.1,2	7		Ontario-E. Lake Superior

<sup>a</sup>The numbers in parentheses are the ISDME (Clark et al. 1987a) computed differences in major axis orientation.



**FIGURE 5.26 ISDME Model Evaluation Region (Clark et al. 1987b)**

geographic area of significant differences was determined by totaling the percentage of half-degree grid cells, within each subregion, where significant differences occur. The results of the analysis showed that ASTRAP significantly overpredicts in all subregions and seasons except the northeast in the summer, northeast and southwest in the spring, and the northeast in the winter. The southeast subregion contained the greatest percentage of regional area for which ASTRAP significantly overpredicts, ranging from 20% in the spring to 54% in the winter and autumn. The subregion with the smallest percentage of regional area with significant overpredictions was the northeast, with less than 29% of its area significantly overpredicted by ASTRAP. No significant overpredictions occurred in this subregion in the spring and winter.

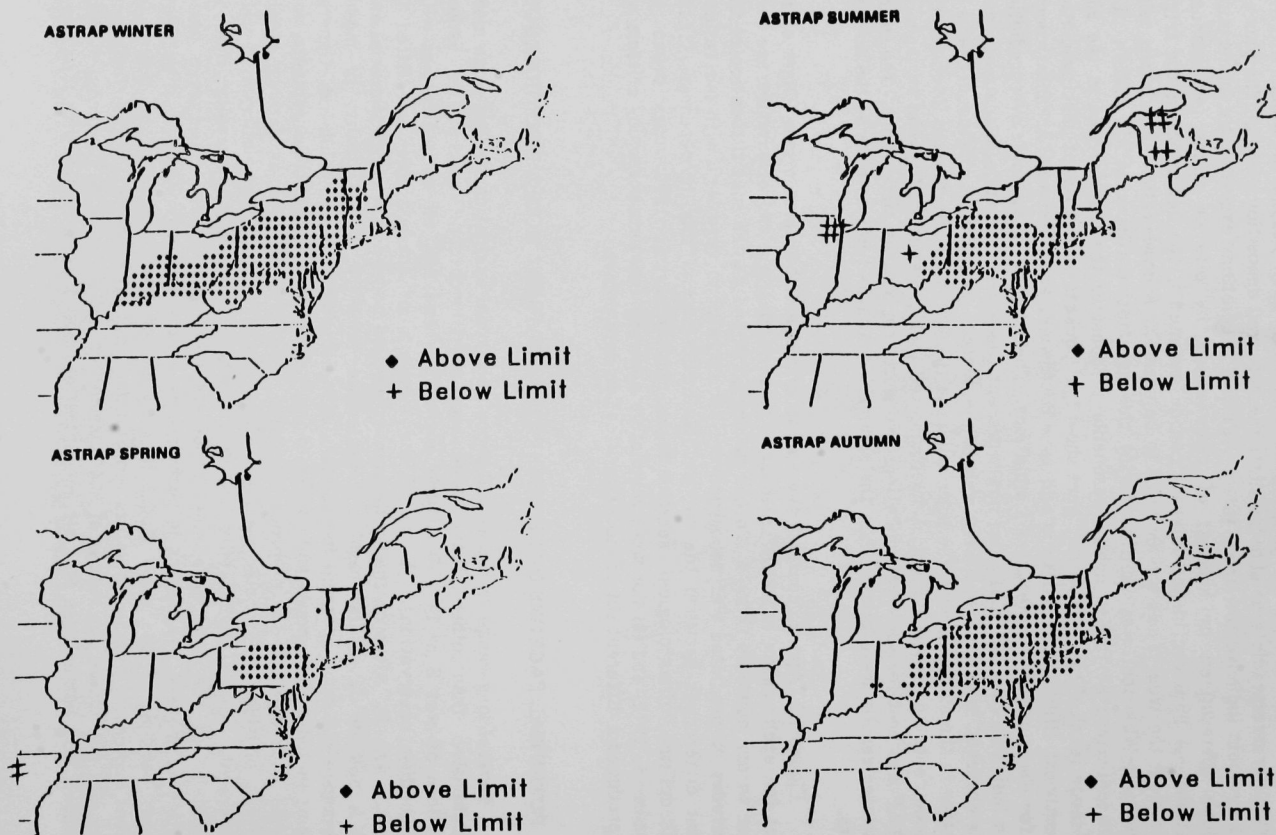


Figure 5.27 shows the geographic areas for each season in which ASTRAP significantly overpredicts or underpredicts the observed deposition values. The plus and minus symbols indicate the areas where ASTRAP predictions exceeded the uncertainty envelope (determined by kriging) of observations by plus or minus two standard deviations. A 95% confidence interval can be assumed if the deviations of the kriged values from the true values are assumed to be normal. A cross-validation analysis using the Shapiro-Wilk test was used to verify this assumption (Clark et al. 1987a). The results indicate that it was not unreasonable to consider the results valid at the 95% confidence level. The data in the figure show that pockets of significant underprediction (in northeast Illinois, central Ohio and New Brunswick, and northwest Arkansas) occur only for the summer and the spring simulations. The extent of geographic overprediction shown indicates that ASTRAP does not represent the spatial variation in the magnitude of the observed spatial pattern very well in the winter and autumn. The model seems to do better in the summer, with the best results in the spring. The explained interseasonal variance reported previously (Sec. 5.3.1) indicated that the explained spatial variance of ASTRAP was best for summer simulations and worst for autumn simulations. The spring and winter results were poorer than the summer results, but not as poor as the autumn results.

Figure I.1, App. I, contains the simple kriged ASTRAP predictions (Clark et al. 1987a) for each season in 1980. Although the general shape and orientation are very similar to our contours (Figs. 5.22b and d, and 5.24b and d), the spatial location of the magnitudes of predicted wet deposition did not correspond very well with our results. Several differences between the two studies could explain this discrepancy: (1) the ISDME-defined climatological season began on January 1, 1980; ours began on December 1, 1979, and (2) the more restrictive observational data screening criteria we used produced a different set of model verification data points.

## 5.4 POTENTIAL FACTORS OF INFLUENCE ON APPARENT MODEL PERFORMANCE

Several long-range transport and deposition model evaluation studies, some more comprehensive than others, have compared daily, monthly, seasonal, or annual predictions of wet S or  $\text{SO}_4^-$  deposition and/or  $\text{SO}_2$  and  $\text{SO}_4^-$  air concentrations with corresponding observations. Some of the elements of these studies are listed in Table B.1, App. B. Although the studies applied similar model performance measures and found a number of similar results, considerable variation in the quality of model performance and some apparent inconsistencies in behavior were also exhibited. General problems included the amount and quality of input data and pollutant measurements, and problems with methods for preprocessing wind field and precipitation data to obtain consistent, physically reasonable results. Different methods for interpolating wind-field measurements yield substantially different wind fields, a phenomenon that is symptomatic of the poor determination of transport winds often cited as a basic limitation of models. Variations in performance among different models and between different versions of the same model in different studies could often be traced to different choices of parameter values representing physical processes, but performance was usually not tested for ranges of such parameters. As a general tendency, models reproduced  $\text{SO}_4^-$  better than  $\text{SO}_2$  air concentrations and worked better with long-term



**FIGURE 5.27** Areas of Significant Differences between ASTRAP Predictions and Observations of Wet Sulfate Deposition in 1980 (♦ or + indicates grid cells of significant overprediction or underprediction) (Clark et al. 1987a)

averages (monthly and annual) than with short-term averages (daily or hourly). Models were not expected to do as well in predicting wet deposition as they did in predicting  $\text{SO}_4^{2-}$  air quality, and in the study conducted by Ruff et al. (1985), this proved to be the case. However, in a study by Stewart et al. (1983) and in the MOI (1982) study, correlations between predictions and observations were higher for wet  $\text{SO}_4^{2-}$  deposition than monthly  $\text{SO}_4^{2-}$  air concentrations. In general, models corresponded relatively poorly with observations in the Ruff et al. study (1985) and MOI (1982) study and substantially better with observations in the Stewart et al. (1983) and Elliassen (1978) studies. A number of explanations have been offered for those variations in results (Ruff et al. 1984), including the fact that different averaging periods were used in the Ruff study, and different data preparation methods were employed. The evaluation of wet-deposition mass flux rather than ionic concentrations and the use of different procedures for adjusting precipitation quantities at monitoring sites have also been suggested as probable causes of the variable results.

These studies, and more recently the NAPAP (1987a,b) Interim Assessment, suggest that some judgments on how well models perform can be made, but few determinations as to why they performed in that manner can be made. This section of our report looks at the "why" in model performance by examining four factors that may account for the noted variations in performance results. The influence of varying model internal parameterization was examined in Sec. 5.2 of this report. The factors that we will examine here are (1) different geographic regions, (2) different sampling protocols, (3) different levels of spatial aggregations of predicted and observed variables, and (4) the use of wet-deposition fluxes versus precipitation-weighted ionic concentrations. The selection of these factors is not intended to suggest that they are the only, or for that matter, the most important influencing factors. Their selection was based primarily on the fact that it is relatively easier to examine their influence than the influence of, for example, alternative model formulations (i.e., the treatment of wind and precipitation fields and the computation of pollutant trajectories).

The influence of different geographic regions on apparent model performance is evaluated by segregating the model evaluation grid into ten geographical regions, identified in Fig. 3.1. Table 5.14 gives the results of model performance across the ten regions as measured by five performance indices. Since the spatial segregation method chosen severely reduced the number of observation/prediction pairs available per region, data from separate seasons were aggregated over individual years and over combined years. Because of the sparsity of valid data points for individual-year aggregation, some regions did not have a sufficient number of observation/prediction pairs to be included in the regional ranking. Regions with N equal to ten or less were not ranked for individual or combined years. Using the Rank Score Index to rank overall performance across regions, ASTRAP performs best in Regions VIII and V and worst in Region IX when all eight seasons are considered. Model performance is best for Regions VIII and V in 1981 and Region VIII in 1980, and worst for Region VII in 1981 and for Region IX in 1980. Therefore, the model seems to perform best in north Ontario (Region V) and the Quebec and Atlantic province (Region VIII).

The model's bias tendency seems to be that ASTRAP overpredicts in high-deposition areas and underpredicts in low-deposition areas, i.e., the model deposition gradient is larger than the observed gradient, particularly in the winter (see deposition

TABLE 5.14 Comparison of Model Performance across Regions

Region	Index of Agreement			Variance Log Error			Dimensionless Mean Square Error			Mean Log Error			Rank Score Index		
	1980 (N) <sup>a</sup>	1981 (N)	1980/81 (N)	1980 (N)	1981 (N)	1980/81 (N)	1980 (N)	1981 (N)	1980/81 (N)	1980 (N)	1981 (N)	1980/81 (N)	1980 (N)	1981 (N)	1980/81 (N)
1	0.86 (10)	0.71 (20)	0.76 (30)	0.055 (10)	0.076 (20)	0.072 (30)	0.165 (10)	0.449 (20)	0.345 (30)	0.182 (10)	0.289 (20)	0.254 (30)	1.57 (10)	2.22 (20)	1.99 (30)
2	0.73 (10)	0.61 (14)	0.67 (24)	0.153 (10)	0.156 (14)	0.156 (24)	0.166 (10)	0.196 (14)	0.183 (24)	-0.345 (10)	-0.272 (14)	-0.302 (24)	2.03 (10)	2.26 (14)	2.13 (24)
3	0.87 (8)	0.59 (14)	0.69 (22)	0.061 (8)	0.179 (14)	0.147 (22)	0.051 (8)	0.189 (14)	0.136 (22)	-0.168 (8)	0.055 (14)	-0.022 (22)	1.44 (8)	2.12 (14)	1.75 (22)
4	- (0)	0.58 (3)	0.58 (3)	- (0)	0.025 (3)	- (0)	0.296 (3)	- (0)	0.296 (3)	0.422 (3)	- (0)	- (0)	2.47 (3)	- (0)	2.47 (3)
5	0.70 (14)	0.32 (18)	0.75 (32)	0.060 (14)	0.126 (18)	0.132 (32)	0.200 (14)	0.099 (18)	0.148 (32)	-0.376 (14)	0.011 (18)	-0.153 (32)	2.06 (14)	1.46 (18)	1.77 (32)
6	0.75 (40)	0.77 (50)	0.75 (90)	0.268 (40)	0.072 (50)	0.173 (90)	0.304 (40)	0.105 (50)	0.197 (90)	-0.439 (40)	-0.206 (50)	-0.309 (90)	2.34 (40)	1.68 (50)	2.01 (90)
7	0.72 (18)	0.66 (19)	0.68 (37)	0.155 (18)	0.224 (19)	0.198 (37)	0.162 (18)	0.179 (19)	0.169 (37)	-0.283 (18)	-0.474 (19)	-0.198 (37)	1.99 (18)	2.39 (19)	2.04 (37)
8	0.91 (11)	0.86 (13)	0.91 (24)	0.035 (11)	0.104 (13)	0.079 (24)	0.092 (11)	0.112 (13)	0.103 (24)	0.223 (11)	0.057 (13)	0.131 (24)	1.45 (11)	1.44 (13)	1.41 (24)
9	0.48 (14)	0.70 (37)	0.63 (51)	0.098 (14)	0.105 (37)	0.108 (51)	0.344 (14)	0.196 (37)	0.240 (51)	-0.547 (14)	-0.388 (37)	-0.431 (51)	3.07 (14)	2.12 (37)	2.37 (51)
10	0.90 (8)	0.62 (7)	0.80 (15)	0.061 (8)	0.092 (7)	0.085 (15)	0.043 (8)	0.095 (7)	0.069 (15)	0.013 (8)	0.205 (7)	0.103 (15)	1.23 (8)	2.00 (7)	1.51 (15)

<sup>a</sup>Numbers in parentheses are the number of observation/prediction pairs (N).

contours in Sec. 5.3). This tendency is indicated by the sign and magnitude of the mean log error in Table 5.14. Underprediction is obvious in Region I (positive MLE) and slight in Regions VIII and X, which tend to be the lower-deposition areas. Overprediction is clear in Regions II, VI, VII, and IX, which tend to be higher-deposition areas. Because Regions VIII, IX, and X tend to span medium- and low- or medium- and high-deposition areas, one should avoid drawing any strong conclusions based on these regions.

Some key factors that may influence the apparent model performance across regions include uncertainties in the spatiotemporal distribution of (1) precipitation amount and rate, (2) emission sources, and (3) transport meteorology. As mentioned earlier, most of these factors cannot easily be investigated. However, differences in sampling protocol (design, sampling period and frequency, chemical analysis, and data interpretation) may also play a role and can be readily investigated. Several years ago, Hakkarinen (1982) compared data from six North American precipitation chemistry networks with daily, weekly, biweekly, and monthly collection frequencies. He found that significant differences in precipitation chemistry measurements can be attributed to differences in sampling protocol. With this in mind, we segregated and grouped data from the seven monitoring networks used in our study into three groups: daily, weekly, and monthly. The model's performance in relationship to these three data groupings is summarized in Table 5.15. The data clearly show a degradation of apparent model performance as sampling frequency increases, with daily or event samplers showing significantly lower model performance levels than the monthly samplers. Performance evaluated with the NADP (weekly) samplers was closer to performance evaluated with monthly samplers than with event samplers. A similar pattern of performance evaluation dependence on sampling protocol exists when the same performance measures are computed for individual seasons (see Table J.1, App. J). If the differences in apparent model performance were a result of sampling error, this pattern of performance with sampling protocol would not be expected. In other words, samples that are left in the field for shorter time periods should be less susceptible to environmental contamination and evaporation.

As noted previously in Sec. 5.1.2, model predictions at MAP3S sites (event samplers) accounted for the greatest number of cases of overprediction by a factor of two (or greater). The seasonal distribution of prediction-to-observation (P/O) ratios that are greater than 2.0 for daily, event, and combined weekly and monthly samplers is given in Table 5.16. Of the P/O ratios at MAP3S sites, 40% or more were greater than 2.0 for all seasons where the factor-of-two overpredictions existed. It was only during the spring and summer of 1981 that no model predictions exceeded observations by a factor of two. The total number of P/O ratios greater than 2.0 at the other daily samplers (none occurred at the APN or APIOS-D network sites) and at the weekly and monthly samplers is much less than the total occurring at the MAP3S sites. The frequency of factor-of-two overpredictions is more than 35% (53 valid samples) at MAP3S sites, with at least one overprediction per site at eight of the nine sites in the MAP3S network. The frequency at which this degree of overprediction occurred is much less (<2%) at the CANSAP and APIOS-C (monthly) sites and NADP (weekly) network sites. More than 75% of the factor-of-two overpredictions can be accounted for when ratios at both event and daily (UAPSP) sites are considered, and more than 94% can be accounted for when event, daily, and weekly (NADP) sites are considered. When the sampling protocol across regions

TABLE 5.15 Protocol Sampling Period Influence on Model Performance

Year	Sampling Protocol <sup>a</sup>	Performance Measures <sup>b</sup>									
		N	IOA	DMSE	VLE	MLE	RSI	MSE	% MSES	SERP (%)	PS
1980	Monthly	38	0.84	0.152	0.157	0.024	1.52	8.5	23.0	17.3	20
	Weekly	37	0.86	0.195	0.157	-0.200	1.72	15.6	25.4	22.9	24
	Daily	63	0.74	0.363	0.151	-0.478	2.34	20.5	68.7	53.2	23
1981	Monthly	54	0.94	0.080	0.090	0.090	1.32	3.0	51.1	29.6	23
	Weekly	86	0.83	0.147	0.147	-0.114	1.61	9.1	43.7	36.6	23
	Daily	70	0.76	0.221	0.171	-0.260	1.97	13.0	40.0	28.7	23
1980/1981	Monthly	92	0.90	0.117	0.119	0.063	1.41	5.3	22.8	17.4	20
	Weekly	123	0.85	0.165	0.152	-0.140	1.63	11.1	21.2	17.5	24
	Daily	133	0.75	0.286	0.174	-0.363	2.16	16.5	58.8	39.3	23

<sup>a</sup>Weekly = NADP; Monthly = APIOS-C, CANSAP; Daily/Event = UAPSP, APIOS-D, MAP3S, APN.

<sup>b</sup>N = number of observation/prediction pairs

IOA = index of agreement

DMSE = dimensionless mean square error

VLE = variance logarithmic error

MLE = mean log error

RSI = rank score index

MSE = mean square error

MSES = systematic mean square error

SERP = systematic error reduction potential

PS = parameter set



**TABLE 5.16 Seasonal Frequency of ASTRAP Factor-of-Two Overpredictions at Model Evaluation Sites as a Function of Sampling Protocol**

Sampling Protocol	1980				1981			
	Win.	Spr.	Sum.	Fall	Win.	Spr.	Sum.	Fall
Event (MAP3S)	3(4) <sup>a</sup>	2(5)	2(5)	4(7)	4(7)	0(7)	0(8)	4(9)
Daily (UAPSP)	2(6)	2(6)	1(6)	2(4)	1(3)	0(6)	0(1)	1(5)
Daily (APN)	0(4)	0(4)	0(4)	0(4)	0(4)	0(4)	0(5)	0(5)
Daily (APIOS-D)	0(0)	0(0)	0(0)	0(3)	0(0)	0(0)	0(9)	0(7)
Weekly/monthly	0(12)	1(12)	0(23)	3(30)	0(22)	0(42)	0(49)	4(44)
Totals	5	5	3	9	5	0	0	9

<sup>a</sup>Numbers in parentheses are the total number of valid samples.

is examined, regions with the poorest overall performance (e.g., RSI >2.1, see Table 5.14) have the greatest number of MAP3S event samplers.\* This suggests that real regional variations in ASTRAP performance, especially in regions with extreme negative bias, may be compounded with apparent bias associated with different observation sampling protocols.

Published studies that compared precipitation chemistry results derived from networks with different sampling protocols were reviewed. The first item to be checked was whether any changes had occurred in MAP3S network operation. The only significant change was in the type of collector. This change was made between November 1979 and April 1981. Before the change, the Battelle Pacific Northwest Laboratory (PNL) design, a polyethylene funnel and bottle precipitation collector, had been employed (Dana and Easter 1987). The current network (after the change) uses the Health and Safety Laboratory (HASL) wet-dry, two-bucket collector. The Aerochem Metrics (ACM) collector, which is patented from the HASL design, is used at most major U.S. networks, including the NADP network. The PNL and HASL collectors were compared in two studies. The first study was conducted during eight months of MAP3S operations. Comparisons of samples taken from co-located PNL and HASL collectors were made at four MAP3S locations (MacCracken 1979). The second study was done at ANL, using samples taken during the summer of 1979 from co-located collectors (Sisterson et al. 1979). The results from both studies showed that the reliability of the two collectors and comparability of the samples were about equal. When these results and the widespread use of ACM collectors (HASL-type samplers) in the NADP network were considered, the

\*The region with the worst overall ASTRAP performance (Region IX) has five of the seven factor-of-two overpredictions from three MAP3S event samplers. Performance in Region II was nearly as bad, with five of six factor-of-two overpredictions at the event samplers.



quality of the samples from different collectors did not seem to be the cause of the large extent of factor of two overprediction at MAP3S sites.

We then looked at studies comparing the precipitation chemistries from samplers with different collection frequencies. Several studies that included comparisons of  $\text{SO}_4^-$  ion concentrations were found (Hakkariinen 1982, Chan et al. 1985, Sisterson and Tisue 1985, Topol and Lev-On 1986; and DePena et al. 1985). Three of these studies were examined in detail: the comparison of North American precipitation chemistry by Hakkariinen, the evaluation of sampling periods and types in Canadian networks by Chan, and a study of data collected over a two-year period by Sisterson, comparing co-located event and weekly samplers at ANL.

Hakkariinen found that networks with longer sampling periods (weekly to monthly) tend to report higher concentrations of  $\text{SO}_4^-$  ions than networks with shorter sampling periods (daily or event); however, no explanation was offered for these differences. Chan found that samples collected at monthly intervals exhibited higher  $\text{SO}_4^-$  concentrations than samples collected daily. The higher  $\text{SO}_4^-$  readings in monthly samples (CANSAP network) were attributed to evaporative losses and contamination from dry deposition. Sisterson used data collected at ANL from April 1980 to March 1982. Chemical differences in the annual and seasonal mean concentrations of nine chemical species were compared. Results showed that the seasonal  $\text{SO}_4^-$  ion concentrations derived from the weekly sampler data were significantly greater (statistically significant differences exceeded analytical uncertainty) than concentrations derived from the event sampler data in five of eight seasons, the exceptions being the summers of 1980 and 1981 and the spring of 1981. The computed concentration differences (weekly-event means) that were significant over the entire period showed that weekly  $\text{SO}_4^-$  ion concentrations exceeded event concentrations by 11.3% in the spring, 22.7% in the fall, and 21.3% in the winter. The reason given for these differences is chemical changes (dissolved  $\text{SO}_2$  oxidation to  $\text{SO}_4^-$ ) that occurred in the weekly sample between the time it was collected and laboratory analysis (Peden and Skowron 1978).

Dissolved  $\text{SO}_2$  [measured as S(IV)] deposited into an event collector has less opportunity for conversion to  $\text{SO}_4^-$  [S(VI)] because the event sample is in the collector for a shorter time and is refrigerated or frozen immediately after collection and kept in this condition until analysis. This conversion can also be minimized by chemically fixing aliquots of the sample with a solution of tetrachloromercurate (TCM) immediately after collection (TCM forms sulfate-complex and aqueous-phase sulfur-IV valence-state  $\text{SO}_3^-$ ) and refrigerating or freezing the sample at or near the same time (Dana 1980). The event samples taken at the ANL site were refrigerated but not chemically fixed. The weekly samples were neither refrigerated, frozen, nor chemically fixed. These weekly samples were preserved by filtering only after arrival at the analytical laboratory, which was from several days to a week after collection. Since cold ambient temperatures in winter would preserve both the event and weekly samples in the field, one might expect to see less significant differences in winter. However, frozen samples were thawed to take an aliquot for pH and conductivity analysis. The event samples were then preserved by refrigeration before further laboratory analysis, while the weekly samples were not (Sisterson et al. 1985). (This procedure is essentially the procedure followed in the MAP3S and NADP networks.) Therefore, the significantly higher  $\text{SO}_4^-$  concentrations in

the weekly samples than the event samples in the winter were most likely the result of conversion of dissolved  $\text{SO}_2$  to  $\text{SO}_4^{2-}$  after collection. Cold-temperature preservation of the event samples in fall and spring 1980 produced similar results. The reason that no significant differences occurred in the summer and spring 1981 samples could be that oxidants (i.e.,  $\text{O}_3$ ,  $\text{H}_2\text{O}_2$ ) were plentiful and the ambient temperatures were warm enough for essentially all of the S(IV) initially in the precipitation to be converted to S(VI) by the time an event sample was collected.\* Dana (1980) found that sulfite in summertime samples at four MAP3S sites was barely detectable. Less sulfite would remain available for gradual conversion in the weekly samples, thus leading to the noted insignificant differences in weekly and event sample  $\text{SO}_4^{2-}$  concentrations in warm seasons.

It appears, therefore, that a major contributing factor to the greater occurrence of significant ASTRAP overpredictions ( $\text{P/O} > 2.0$ ) with event or daily collectors than with weekly or monthly collectors, particularly during the colder seasons, is the more complete oxidation of S(IV) to S(VI) for collectors on longer sampling protocols. The cold temperatures and absence of oxidants in the winter, with a resulting observed S(IV) maximum in the winter (Dana 1980), would seem to explain the observed  $\text{SO}_4^{2-}$  deposition minimum in samples that are preserved. The S(IV) would gradually be converted to S(VI) in samples that are not preserved. The wet-removal parameterization in ASTRAP is for bulk sulfur (i.e., removal rates for  $\text{SO}_2$  and  $\text{SO}_4^{2-}$  are identical). While initial wet removal of  $\text{SO}_4^{2-}$  in the atmosphere is more efficient than removal of  $\text{SO}_2$  (the  $\text{SO}_4^{2-}$  aerosol serves as cloud condensation nuclei), in-cloud oxidation of  $\text{SO}_2$  can be rapid, especially in the summer when oxidants are plentiful and total sulfur deposition is several times larger than observed in the winter. The wet sulfur deposition predicted by ASTRAP corresponds to the bulk sulfur equivalent of combined S(IV) and S(VI) and should, if S(IV) is not measured, more closely correspond to observations in which sample preservation of S(IV) is not ensured (NADP, APIOS-C, CANSAP, etc.). This, in fact, seems to be the case, as indicated with the results presented in Table 5.15.

The third factor investigated that could potentially influence model performance is a change in spatial scale for averaging paired observations and predictions. The scales selected for these pairings are individual site, unit-grid increments (30-40 km); nine-grid increments (300-390 km); and twelve-grid increments (1,200-1,560 km). Figure 3.1 shows the relative sizes of these grid scales. The spatial average deposition within each grid was computed as a simple arithmetic average. Table 5.17 summarizes the statistical measures used to examine changes in model performance over the four chosen levels of spatial aggregation. Grouped seasonal data were used for each year (1980 and 1981) and for the combined years. The comparisons show that model performance improves as the

---

\*Hales and Dana (1979) found that the solubility of  $\text{SO}_2$  increases with increasing temperature, decreasing free acidity (increasing pH), and increasing gas-phase  $\text{SO}_2$  concentrations. In addition, laboratory kinetic studies and atmospheric measurements have shown that  $\text{H}_2\text{O}_2$  may be the key atmospheric oxidant of dissolved S(IV) species ( $\text{SO}_2$ ,  $\text{SO}_3^{2-}$ ,  $\text{HSO}_3^-$ ) at solution pH  $< 5.0$  (Schwartz 1984, Lee et al. 1986). This is due to the high aqueous solubility of  $\text{H}_2\text{O}_2$  and its increase in reaction rate with S(IV) as acidity increases (Kleindiest et al. 1988).

**TABLE 5.17 Comparison of Model Performance Based on Spatial Aggregation**

Year	Level of Aggregation	Performance Measures <sup>a</sup>									
		N	IOA	DMSE	VLE	MLE	RSI	MSE	% MSES	SERP (%)	PS
1980	PI <sup>b</sup>	139	0.803	0.253	0.199	-0.263	1.96	15.7	29.4	13.4	17
	Unit	133	0.795	0.262	0.203	-0.269	1.99	16.2	29.6	13.4	17
	9	77	0.821	0.209	0.150	-0.222	1.80	11.3	26.4	15.9	17
	36	34	0.860	0.144	0.089	-0.203	1.60	6.8	32.0	25.8	17
1981	PI	230	0.836	0.158	0.155	-0.108	1.62	8.7	17.1	0.3	11
	Unit	195	0.837	0.158	0.163	-0.105	1.62	8.2	15.4	0.5	11
	9	102	0.844	0.152	0.153	-0.066	1.56	6.2	7.9	0.3	11
	36	38	0.896	0.087	0.114	-0.034	1.35	3.2	5.2	0.5	11
1980/ 1981	PI	369	0.820	0.197	0.177	-0.166	1.76	11.4	20.3	4.7	18
	Unit	328	0.815	0.205	0.186	-0.171	1.79	11.4	20.0	7.1	18
	9	179	0.830	0.181	0.158	-0.133	1.68	8.4	14.6	8.4	18
	36	72	0.877	0.118	0.109	-0.114	1.48	4.9	15.3	12.5	18

<sup>a</sup>N = number of observation-prediction pairs

IOA = index of agreement

DMSE = dimensionless mean square error

VLE = variance logarithmic error

MLE = mean log error

RSI = rank score index

MSE = mean square error

MSES = systematic mean square error

SERP = systematic mean square error

PS = parameter set

<sup>b</sup>PI = paired individually

spatial aggregation scale gets larger. This may be because as the aggregation scale increases, the influence of extreme values (with the larger number of data points available) is likely to be reduced. There is essentially no difference in model performance between the pairing of observations and predictions on an individual-site or a single-grid basis, since few grid cells have more than a single sampling site. The degree of improvement in model performance in aggregation over larger scales is greater for the 1980 data than the 1981 data.

The last factor of influence in model performance that was investigated is the use of PWICs. The seasonal  $\text{SO}_4^{=}$  concentrations reported in the Acid Deposition System (ADS) data base included  $\text{SO}_4^{=}$  computed as a PWIC (in mg/L) and deposition computed from PWIC as a wet sulfur mass surface flux ( $\text{g/m}^2$ ). A description of the computations in ADS for deriving PWIC and deposition is given in App. K. Table 5.18 summarizes the statistical measures used to compare model performance when observations and predictions are paired as  $\text{SO}_4^{=}$  PWIC versus mass deposition flux. Eight seasons of data are reported. The data show that model performance for four simulated seasons (winter, spring, and summer 1980, and autumn 1981) declined when observations and predictions were compared on a PWIC basis. The opposite was true for the winter 1981 simulations (RSI goes from 2.20 to 1.99). The simulations for the other seasons did not show any significant change in performance with PWIC versus mass flux. Significant declines in the explained spatial variance and total EBCV occurred for all seasons when data were compared on a PWIC basis versus a mass flux basis (the changes are given in Table 5.17). These results are somewhat surprising, since one might assume that precipitation weighting would smooth out some of the spatial error inherent in collecting and spatially interpolating precipitation data. The decline in the model's ability to explain variance with use of PWIC may be due to the fact that ASTRAP calculations utilize grided precipitation fields rather than values measured at the wet-deposition observation sites. Venkatram et al. (1986) and others have found that when receptor-specific precipitation data were substituted into the diagnosed precipitation field, the agreement between observed and predicted  $\text{SO}_4^{=}$  PWIC improved substantially. Although these results were obtained with an episodic model, similar results were obtained by Clark et al. (1987a) with a Lagrangian model for seasonal predictions.

**TABLE 5.18 Comparison of Model Performance Based on Precipitation-Weighted Ionic Concentration Versus Mass Flux**

Season	Form	Performance Measures <sup>a</sup>						TVE (%)	SVE (%)	R <sup>2</sup> (%)
		IOA	DMSE	VLE	MLE	RSI				
Winter 80	PWIC <sup>b</sup>	0.41	0.30	0.221	-0.197	3.16				
	Mass Flux	0.49	0.27	0.262	-0.246	2.82				
Winter 81	PWIC	0.72	0.23	0.137	-0.232	1.99				
	Mass Flux	0.67	0.27	0.242	-0.199	2.20				
Winters	PWIC						38.1	-10.8	27.3	
	Mass Flux						39.7	12.3	52.0	
Spring 80	PWIC	0.55	0.28	0.121	-0.294	2.51				
	Mass Flux	0.63	0.24	0.121	-0.351	2.29				
Spring 81	PWIC	0.77	0.13	0.131	-0.039	1.60				
	Mass Flux	0.82	0.11	0.116	0.015	1.46				
Springs	PWIC						33.3	-5.4	27.9	
	Mass Flux						32.8	16.5	49.5	
Summer 80	PWIC	0.76	0.24	0.136	-0.030	1.72				
	Mass Flux	0.85	0.15	0.138	-0.055	1.52				
Summer 81	PWIC	0.85	0.10	0.084	0.120	1.48				
	Mass Flux	0.85	0.11	0.100	0.054	1.43				
Summers	PWIC						29.5	25.7	55.2	
	Mass Flux						34.4	41.2	75.6	
Autumn 80	PWIC	0.56	0.44	0.211	-0.325	2.76				
	Mass Flux	0.57	0.40	0.215	-0.410	2.78				
Autumn 81	PWIC	0.51	0.39	0.132	-0.359	2.84				
	Mass Flux	0.70	0.23	0.135	-0.315	2.11				
Autumns	PWIC						24.6	-49.9	0	
	Mass Flux						36.5	4.0	40.5	

<sup>a</sup>IOA = index of agreement; DMSE = dimensionless mean square error;  
VLE = variance logarithmic error; MLE = mean logarithmic error;  
RSI = rank score index; TVE = temporal variance explained;  
SVE = spatial variance explained; R<sup>2</sup> = residual correlation coefficient

<sup>b</sup>PWIC = precipitation-weighted ionic concentration

## 6 CONCLUSIONS AND RECOMMENDATIONS

After previous model evaluation studies (App. B) were completed, several questions remained about the reliability and performance of Lagrangian meso- and synoptic-scale transmission\* and deposition (LMSTD) models.† The nature of these technical issues, in turn, raised important questions about the ability and ultimate utility of these models for formulating policies on controlling acid deposition. Some of the key model performance issues raised are summarized here:

- Can temporal patterns in  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations and in wet  $\text{SO}_4^-$  deposition be reasonably reproduced?
- How well can the magnitudes of maximum seasonal  $\text{SO}_4^-$  depositions and  $\text{SO}_4^-$  and  $\text{SO}_2$  air concentrations (DAC) be reproduced?
- Are there significant differences in interannual seasonal (e.g., between summers of different years) and interseasonal (e.g., between winter and summer of the same year) performance of LMSTD models?
- How sensitive is model performance to linear adjustments in the empirical parameterization of LMSTD models?
- Are there significant interannual variations in the magnitudes of predicted and observed wet deposition because of variability in meteorology? Does this variability severely restrict our ability to forecast reliable deposition patterns?
- How well can we reproduce the spatial patterns in DAC observations; i.e., can the spatial patterns, such as DAC contour orientation, shape, and gradient magnitude, be reasonably reproduced?
- How well do LMSTD models perform in different geographical regions, and can the basis for observed significant differences in performance be reasonably identified?
- Can we decompose error components into spatial, temporal, and bias parts or into systematic or unsystematic parts to gain a greater insight into model performance?

---

\*Transmission = transport and diffusion, chemical transformation, and wet and dry removal (precipitation scavenging, dry-deposition physics, etc.).

†For use in predicting long-term (one month or longer) air-pollutant concentration averages and pollutant-deposition flux accumulation.



The success of providing a reasonable reply to these issues is contingent upon (1) the spatial resolution of the available model evaluation data base, (2) our knowledge of the magnitude of error in this data base, and (3) the development of a model evaluation methodology capable of bringing out the salient features of spatiotemporal and bias error patterns. Although we feel the observational data base (DAC) used in our study is adequate to evaluate the LMSTD model, we lack the data necessary to quantify the error in this data base. The same lack of data on meteorology and emissions data (the input to the model) exists. This lack of knowledge may preclude us from providing complete and definitive responses to the above-referenced issues and to our study objectives. (Quantification of error in the data base is an issue addressed in our recommendations.)

The model evaluation methodology was selected to provide a group of performance measures that could collectively meet our model evaluation objectives and address the aforementioned model performance issues. Because no single or narrow group of measures can be expected to provide a comprehensive response or a response in which reasonable confidence can be placed, we borrowed, developed, evaluated, tested, and adapted a broad group of statistical tools to meet our needs. These included statistical spatiotemporal pattern display methods and parametric statistics for quantifying residuals. The pattern display methods chosen and developed included scatter and time-series plots, residual and observation-prediction histograms, normalized and fractional bias and scatter error plots, and spatial pattern trends analyses or drift estimation techniques (spatial interpolation methods such as simple and universal kriging). A large group of descriptive statistics supplemented the information provided by the graphical methods. Some of the more important measures included mean square error (MSE) and its dimensionless derivatives (i.e., spatial, temporal, bias, systematic, and unsystematic components), index of agreement (IOA), rank score index (RSI), relative DMSE (also a MSE derivative), bias, variance, correlation, and a number of logarithmic indices.

## 6.1 SUMMARY OF MAJOR FINDINGS

Our results can be categorized under four headings: residual and scatter error patterns, performance sensitivity in apparent model error, spatiotemporal and bias error patterns, and potential factors of influence on apparent model performance. The key findings under these four categories are given here. Greater physical understanding of model performance could be gained if predictions of wet deposition and average air concentrations were evaluated for the same periods. For instance, simulated atmospheric concentrations might be too low because parameterized wet removal is too high, but if simulated wet deposition for the same period is also too low, then some other feature must be involved. Unfortunately, suitable observation data sets for wet deposition and regional air quality did not coincide. References to the model performance issues raised in the introduction to this section are made when appropriate.



### 6.1.1 Residual and Scatter Error Patterns

The residual and scatter error in ASTRAP predictions are determined by applying parametric statistical measures to quantify the magnitude of the apparent model error and using graphical statistics to visually display the patterns of this error. The results indicate that the magnitudes of maximum seasonal and monthly mean DAC can be reproduced reasonably well. Without additional years of model evaluation data, however, it is not possible to determine whether temporal patterns in observed wet  $\text{SO}_4^-$  deposition are also reasonably reproduced by the model. The limited data analyzed show significant interseasonal and interannual (among winters, springs, and autumns) differences in model performance. More data are needed to confirm this. The major findings on ASTRAP's performance in simulating monthly average air concentrations and then in simulating seasonal wet deposition are highlighted.

#### 6.1.1.1 Monthly Air Concentrations

- Residuals for monthly (January, April, and July)  $\text{SO}_4^-$  air concentration are almost normally distributed, with at least 60% of the differences in observations and predictions within one standard deviation of the mean residual ( $\sigma_r$ ). The October distribution is skewed to the left (strong tendency toward overprediction), with only 30% of the residuals within one  $\sigma_r$ . A slight to moderate positive bias exists in monthly  $\text{SO}_4^-$  predictions for July and January, and a slight negative bias exists for April.
- The model overpredicts observed  $\text{SO}_4^-$  concentrations by greater than a factor of two at one location in January and at five locations in October. Model underpredictions of the same degree occur at two locations in January and one location in July.
- The residual distributions for monthly  $\text{SO}_2$  air concentrations approach normality for all four simulation months. A positive bias (underprediction) tendency in  $\text{SO}_2$  predictions is exhibited for all simulation months, more so for January and April than for July and October.
- The model underestimates observed monthly  $\text{SO}_2$  air concentrations by greater than a factor of two for 12 of 145 observations: three in January and July, four in April, and one in October. The same degree of overprediction occurred only at single locations in January and July.
- Although the bias error is smaller over the first three months for  $\text{SO}_4^-$ , ASTRAP seems to track the variations in mean monthly  $\text{SO}_2$  observations better than the variations in mean monthly  $\text{SO}_4^-$  observations. Local variations in meteorology and emissions, not adequately resolved in our model input data base, may be the major

contributing factor leading to this underprediction in  $\text{SO}_2$  simulations.

- The relative error (expressed as the DMSE percentage of the  $\text{MSO}/(\text{O})^2$  in ASTRAP monthly  $\text{SO}_4^-$  predictions ranged from 4% (July) to 33% (October), while the relative error for ASTRAP monthly  $\text{SO}_2$  predictions ranged from 18% (October) to 25% (January).
- ASTRAP is able to explain 49% of the variance in July  $\text{SO}_4^-$  observations and 37% of the variance in January and July  $\text{SO}_2$  observations. Only 15% of the January observed  $\text{SO}_4^-$  variance and 26% of the April observed  $\text{SO}_2$  variance is explained.
- As measured by a robust performance statistic, RSI, ASTRAP does very well in simulating July  $\text{SO}_4^-$  air concentrations. Performance is adequate when simulating July  $\text{SO}_2$ . The model performs poorly when predicting  $\text{SO}_4^-$  and  $\text{SO}_2$  in January and April, and when predicting  $\text{SO}_4^-$  in October. Less than adequate performance is exhibited when simulating October  $\text{SO}_4^-$ . The very good model performance for July  $\text{SO}_4^-$  may be due, in part, to the availability of better empirical data used for ASTRAP parameterization of summer transformation rates ( $\text{SO}_2$  to  $\text{SO}_4^-$ ) than the data available for other seasons. ASTRAP's predictions of  $\text{SO}_4^-$  concentrations are most sensitive to variations in transformation rates (see discussion in Sec. 6.1.2).

#### 6.1.1.2 Seasonal Wet Sulfate Deposition

- The summer 1980 and 1981 residuals for wet  $\text{SO}_4^-$  seasonal deposition are approximately normally distributed, with over 70% of the observation-prediction differences within one  $\sigma_r$ . The residuals for winter 1980 show the farthest departure from normality, with only 44% of the residuals within one  $\sigma_r$ . When seasons are combined, the 1981 distributions of residuals are shown to approach normality (76% within one  $\sigma_r$ ), and the 1980 distributions are shown to depart from normality (only 47% within one  $\sigma_r$ ). The greater number of data points (observation-prediction pairs) available in 1981 probably caused the difference in the 1980 and 1981 distributions.
- Model predictions for summer 1980 and spring and summer 1981 show the smallest scatter (with respect to other seasons) and the most symmetrical (unbiased) scatter around the perfect prediction line. Comparisons of observations and predictions for the remaining seasons show a tendency for model overprediction.

- The model overpredicts observed  $\text{SO}_4^-$  deposition by more than a factor of two for 22 of 133 observations in 1980 and 14 of 195 observations in 1981. Over 75% of these overpredictions occurred in the winter or the fall at sites on an event or daily sampling protocol. The same degree of underprediction occurred at only three locations, one in summer 1981 and two in winter 1981.
- It is interesting to note that the factor-of-two over- and underprediction data points are predominately represented by precipitation chemistry samplers with event (MAP3S) and daily (UAPSP) sampling protocols. This suggests that there may be systematic differences in observations resulting from sampling protocol (see Sec. 6.1.4).
- The time series of the apparent bias in predictions shows a stronger tendency toward model overprediction in 1980 than 1981, especially in the spring. The ratio of the coefficient or variation of predictions to the coefficient or variation of observations (CVR) is close to 1.0 for all seasons except winter 1980 (CVR = 0.52), winter 1981 (CVR = 0.84), and spring 1980 (CVR = 1.8). The model seems to track the season-to-season variation in observed wet  $\text{SO}_4^-$  deposition fairly well.
- The relative error [expressed as the DMSE percentage of the  $\text{MSO}/(\text{O})^2$ ] in ASTRAP seasonal wet  $\text{SO}_4^-$  deposition predictions ranged from 9% (summer 1981) to 34% (fall 1980).
- Model simulations in summer and in spring and fall 1981 exhibit about 50% to 55% of the observed variance, while simulations in winter 1980 exhibit only 5% of the observed variance. The explained bias-corrected variance (EBCV) between seasons, decomposed into spatial and temporal components, is given in Sec. 6.1.3.
- Using a robust statistical measure, RSI, we found that ASTRAP performs best in simulating seasonal  $\text{SO}_4^-$  wet deposition for summer and spring 1981. Performance drops off sharply when simulating seasonal observations in autumn and winter 1980. The interannual differences in model performance appear significant for all seasons, except winter and fall of 1980 and spring and summer of 1981. These results should be viewed with caution because of the limited amount of data available (only 2 years) for comparing differences in interannual and interseasonal model performance.

### 6.1.2 Performance Sensitivity Patterns In Apparent Model Error

The sensitivity in ASTRAP model performance is examined by a factor-of-two adjustment to four internal model parameters ( $V_d$  for  $\text{SO}_2$  and  $\text{SO}_4^-$ ,  $T_p$ , and WC). The

dry-deposition velocities ( $V_d$ ) were varied in the same direction simultaneously and counted as a two-parameter variation. All references to ASTRAP pertain to results for the unmodified version. Since variation of internal model parameters does not greatly alter the shape, position, and orientation of the predicted contours, the performance sensitivity pattern analysis did not consider the model's sensitivity in reproducing observed spatial patterns. The results indicate the uncertainty variations (100% larger or 50% smaller) in ASTRAP parameters considered in this study can significantly influence the magnitude of model predictions and the level of performance in those predictions. Model estimates of  $\text{SO}_2$  and  $\text{SO}_4^{=}$  air concentrations are most sensitive to variations in  $V_d$  and transformation rate ( $T_r$ ), respectively. Overall, the predicted  $\text{SO}_4^{=}$  air concentrations are more sensitive to parameter variations than the predicted  $\text{SO}_2$  air concentrations. Model estimates of wet deposition are most sensitive to variations in wet-removal coefficient (WC). The major specific findings are highlighted in the following sections.

#### 6.1.2.1 Monthly Air Concentrations

- Distinct performance sensitivity patterns emerge for  $\text{SO}_4^{=}$  and  $\text{SO}_2$  simulations when the selected model parameter adjustments are made. The performances of the 27 model versions resulting from these adjustments (including the nominal or standard ASTRAP) are clustered on a normalized bias-scatter error plot in parameter sensitivity groups of three (triple) and nine (triplet). The triple clustering for  $\text{SO}_2$  is performance-ordered by  $T_r$ , while the triple clustering for  $\text{SO}_4^{=}$  is performance-ordered by  $V_d$ . The triplet clustering for  $\text{SO}_2$  is performance-ordered by  $V_d$ , while the triplet clustering for  $\text{SO}_4^{=}$  is performance-ordered by  $T_r$ . The individual model versions within triples are performance-ordered by WC. There is less performance sensitivity for  $\text{SO}_2$  triples and triplets than for  $\text{SO}_4^{=}$  triples and triplets. In other words, internal parameter variations affect model performance more when  $\text{SO}_4^{=}$  air concentrations are simulated than when  $\text{SO}_2$  air concentrations are simulated.
- Simulations of  $\text{SO}_2$  are most sensitive to variations in  $V_d$ , followed by variations in  $T_r$  and WC. Simulations of  $\text{SO}_4^{=}$  are most sensitive to variations in  $T_r$ , followed by  $V_d$  and WC. Because  $\text{SO}_4^{=}$  concentrations are linear functions of  $\text{SO}_2$  emissions in ASTRAP but are affected by  $T_r$  in an opposite way than are  $\text{SO}_2$  concentrations, the extremes produced by parameter adjustments require that parameters with the most sensitive influence on air concentrations ( $V_d$  and  $T_r$ ) be adjusted in opposite directions.
- By making the parameter adjustments considered, the model's ability to reproduce observed monthly  $\text{SO}_4^{=}$  air concentrations improved by as much as 34% (by halving the  $V_d$  and  $T_r$  and doubling the WC for October predictions). No significant improvements were evident in July  $\text{SO}_4^{=}$  and  $\text{SO}_2$  and October  $\text{SO}_2$  predictions.

- The systematic error reduction potential (SERP), through internal model parameter adjustment, is greatest for October  $\text{SO}_4^-$  simulations (SERP = 59%). The SERP for  $\text{SO}_2$  and  $\text{SO}_4^-$  predictions in April and January is around 23%. This suggests that a significant fraction of the systematic error in model predictions for these particular monthly periods can be reduced through adjustments in model parameterization. The systematic error in model predictions is inherent not only to model parameterization but also to the measurement of air concentrations, quantification of source emissions, and generation of wind and precipitation fields. Without a means to segregate the sources of systematic error in model predictions, any revisions to model parameterization should be made cautiously, and in any case, only with the support of data obtained through field verification. Recommendations are made in Sec. 6.2 on the importance of quantifying model input and evaluation data (field measurements of DAC) error and on the statistical treatment of this error to more readily identify its sources (i.e., model parameterization, field measurements).

#### 6.1.2.2 Seasonal Wet Sulfate Deposition

- Distinct patterns in model performance sensitivity emerge in the wet-deposition simulations when the selected model parameterization adjustments are made. The performances in model predictions resulting from these adjustments are clustered on normalized bias-scatter error plots in groups of three (triple) and nine (triplet). Triple clustering is performance-ordered by  $V_d$ , while triplet clustering is performance-ordered by WC. For the parameter adjustments considered, these patterns reveal that model performance for wet  $\text{SO}_4^-$  deposition is most sensitive to variations in WC, followed by variations in  $V_d$ .
- The parameter adjustments did not improve the model's ability to reproduce wet deposition as much as they had for air concentrations. A maximum of 16% improvement over ASTRAP's performance occurred for the autumn simulations, with WC halved and with  $T_r$  normal or doubled. None of the parameter-set adjustments considered provided better results than the results obtained with ASTRAP summer and spring 1981 simulations.
- The SERP achievable through our internal model parameter adjustment is greatest for autumn (41% in 1981 and 14% in 1980) and spring (22% in 1980) wet  $\text{SO}_4^-$  deposition predictions. These results suggests that a significant fraction of the systematic error in model predictions for these seasonal periods, most notably autumn 1981, could be reduced through adjustments in model parameterization.

- Since systematic error is inherent to model input and model evaluation data bases, any revisions to model parameterization should be made cautiously and should be based upon relevant field measurements. Recommendations are made in Sec. 6.2 on the importance of quantifying model input and model evaluation data (field measurements of wet  $\text{SO}_4^-$  deposition) error, and a method is suggested to treat this error statistically in order to identify its sources more readily.

### 6.1.3 Spatial, Temporal, and Bias Error Patterns

Error patterns are examined by decomposition of MSE into spatial, temporal, and bias components and by decomposition of variance into spatial and temporal components. Kriging is then used to further examine ASTRAP's ability to reproduce spatial patterns in the observation fields (such as the position, shape, orientation, and magnitude of the gradient in the contours). Although there are no significant variations in the magnitudes of the maximum predicted and observed interannual wet depositions, there are significant variations in the locations of the observed maxima. Since variability in meteorology plays a substantial role in influencing locations where observed maxima occur, the way wind and precipitation fields are treated (in most Lagrangian and in some Eulerian models) may be the reason the model has trouble locating the maximum deposition areas. All regional transport models have difficulty properly characterizing local and subgrid variations in wind and precipitation fields. The model's inability to reproduce other spatial features in the observed data, such as the shape, orientation, and magnitude of the deposition gradient, may also result from problems in representing wind and precipitation fields, although simplifications inherent in parameterizations of chemical or removal processes may also contribute. With respect to wind fields, Kuo et al. (1984) illustrated improvements in trajectory accuracy, versus the use of normal NMC observations, through use of numerical weather prediction (NWP) models to generate winds. These models would, in effect, be used as a sophisticated spatiotemporal methodology for both winds and thermodynamic atmospheric properties. The winds would be mass-consistent, dynamically correct, more reliable in data-sparse areas, and potentially more descriptive of phenomena such as the nocturnal jet (Demerjian 1985). The specific major findings follow:

- The spatial error in ASTRAP predictions of wet  $\text{SO}_4^-$  deposition dominates, accounting for over 70% of the total error in the winter, spring, and summer. The predicted wet  $\text{SO}_4^-$  deposition in the autumn shows comparative levels of spatial and bias error, with a relatively small contribution of temporal error to the total error. The temporal error across seasons is smaller than the other two error components, particularly for winter and autumn. These results are probably caused by the statistically small number of data points (two seasons) considered in our analysis. Although only 45% of the error in the autumn predictions is spatial in origin, the relative larger overall error in autumn (73% to 122% greater than in spring and summer) makes the spatial error in autumn slightly larger than that in spring and summer.



- The model's ability to explain bias-corrected variance from one year to another ranges from 41% for autumn simulations to over 75% for summer simulations. The percentage of the total EBCV resulting from the model's ability (inability) to reproduce the spatial patterns in observations ranges from 10% for autumn simulations to 54% for summer simulations. In contrast to this rather considerable seasonal variation, a relatively small seasonal variation is shown for the explained temporal variance. A maximum of only 7% variation across seasons is exhibited for the temporal variance, while over 35% variation is exhibited for spatial variance. The model appears to explain temporal variance from year to year well, but it does not explain spatial variance from year to year so well, with the exception of the summer simulations. Although these results seem to be in line with results obtained through MSE decomposition, additional years of data are needed to confirm these findings.
- A geostatistical interpolation technique known as kriging was used to visually assess ASTRAP's ability to reproduce spatial patterns in DAC. This analysis showed that although the magnitudes of the observed maximum in  $\text{SO}_4^-$  air concentrations and wet  $\text{SO}_4^-$  deposition are reasonably reproduced, the locations of these maximums is not. The model also had difficulty in reproducing the position, shape, and magnitude of the gradient in the observed spatial patterns.

#### 6.1.4 Potential Factors of Influence on Apparent Model Performance

Four factors that may account for some of the performance results for wet  $\text{SO}_4^-$  deposition obtained in this study and for the variability in performance results obtained in previous model evaluation studies were examined. The factors include (1) geographic regions, (2) sampling protocols, (3) levels of spatial aggregation of predicted and observed variables, and (4) the expressed forms of the observed and predicted variable (wet deposited fluxes versus precipitation-weighted ionic concentrations). The major findings follow.

- ASTRAP seems to perform substantially differently in different geographic regions. Real regional variations in model performance, especially in regions with extreme negative bias, may be exacerbated by the apparent bias associated with observation sampling protocol (see next item).
- Significant ASTRAP overpredictions (by more than a factor of two) occurred much more often at observation sites with an event or daily sampling protocol than at sites with a weekly or monthly sampling protocol, particularly during cold or cool seasons. This appears to be the result of the fact that samples with sulfur(IV) are preserved before they are analyzed at sites with an event or daily



sampling protocol (i.e., MAP3S). The sulfur(IV) in samples that are not preserved (i.e., with a weekly or monthly sampling protocol) gradually converts to  $\text{SO}_4^{2-}$  [sulfur(VI)], thus leading to higher  $\text{SO}_4^{2-}$  measurements in samplers with these protocols. The wet sulfur deposition in ASTRAP corresponds to the bulk sulfur equivalent of combined S(IV) and S(VI) and should more closely correspond to observations in which sample preservation of S(IV) is not ensured.

- Aggregation and averaging of site observations and predictions over larger spatial scales before comparison showed that model performance improved as the aggregation scale got larger. This is probably a result of smoothing of smaller-scale variations in observations associated with terrain effects on mesoscale meteorology or with local sources.
- When observations and predictions are compared on a basis of PWIC versus mass deposition flux, ASTRAP performance declined for simulations in four seasons, did not show any significant change for three seasons, and increased for one season. These results may have occurred because ASTRAP calculations use gridded precipitation fields, not values measured at the wet-deposition sites.

## 6.2 RECOMMENDATIONS

Although this study helps to provide new insights on model performance evaluation (MPE) methods and a better understanding of MPE results, we are still unable to specify the level of uncertainty in model predictions and we still lack a fundamental understanding of why LMSTD models perform the way they do. There are three areas of further research could result in a way to quantify uncertainty and improve our understanding of model performance. These research areas are (1) the completion of the development and test application of the empirical Bayesian uncertainty quantification methodology, (2) the estimation of model input and model evaluation data errors and the explicit incorporation of these errors into measures of performance of model predictions, and (3) the investigation of the sensitivity in model performance by local and global variation of key model and model input variables. This recommended research is briefly described.

### 6.2.1 Empirical Bayesian Probability Methodology

To explicitly quantify uncertainty in model predictions, we must make hypotheses about that uncertainty and then test them. As previously mentioned in this report, this requirement can be accomplished through application of a Bayesian

probability formulation.\* Bayesian methods, however, are often criticized because they introduce heuristics or subjectivism into the process. Instead of depending on subjective judgment to derive the prior distribution of model-predicted variances (prior to comparison of predictions with current observations), a method called an empirical Bayesian approach (EBA) can be used to derive the prior distribution through use of (for lack of a better term) a baseline observational data base. We have already developed a modified form of the EBA that assumes the prior distribution, but key parameters in that distribution are derived empirically.

Our principal interest in the EBA is that it provides a means to compute the probability of the outcome (success or failure) of a set of proposed actions (i.e., policy options for the control or mitigation of acid deposition) based on the computed uncertainty in the predicted variables that are pertinent to judging the outcome of policy options. The probability of outcome of the proposed actions is expressed in terms of the collective uncertainty of variables directly or indirectly affecting the outcome. Uncertainty, applied to a model, generally pertains to the range of expected error between the model predictions of a quantity and actual values of that quantity. Since the model ordinarily predicts a number of values with varying errors, both error and uncertainty must be described in terms of distributions of values. We seek various measures of those distributions, including means, standard deviations, differences among various parts of the model domain, and ideally, a representation of the distribution itself. For example, uncertainty could be quantitatively expressed in terms of the joint conditional probability distribution for a set of true values  $x_1, x_2, \dots, x_n$ , given the corresponding model predictions  $x_1, x_2, \dots, x_n$ . The task of an EBA is to estimate that distribution, or at least certain measures of it, based in part on samples of apparent error. However, it should be emphasized that uncertainty distributions reflect our knowledge of the model performance and thus are subject to change as increased information becomes available. For example, we might expect that the MSE between a given set of predictions and true values would become better defined (i.e., smaller variance about the expected value) as more data become available for evaluation of the model.

---

\*"In the Bayesian approach to statistics, an attempt is made to utilize all available information in order to reduce the amount of uncertainty present in an inferential or decision-making problem. As new information is obtained, it is combined with any previous information to form the basis for statistical procedures. The formal mechanism used to combine the new information with the previously available information is known as Bayes' theorem; this explains why the term 'Bayesian' is often used to describe this general approach to statistics. Bayes' theorem involves the use of probabilities, which is only natural, since probability can be thought of as the mathematical language of uncertainty. At any given point in time, the statistician's (or the decision maker's) state of information about some uncertain quantity can be represented by a set of probabilities. When new information is obtained, these probabilities are revised in order that they may represent all of the available information." (Winkler 1972)

Figure 6.1 shows some of the major components of an empirical Bayesian analysis framework for use in generating uncertainty distributions. A great deal of effort went into developing and testing the empirical Bayesian model (EBM) during the initial stages of this project. Additional work is needed to modify the "Bayes" integrals, used for computing prior and posterior distributions and spatially averaged moments (means and variances) to account for quantifiable biases inherent to the data base. Upon consideration of the effort already expended, the availability of four additional years of precipitation chemistry data (through 1986), and the important role source-receptor uncertainty analysis could play in the future analysis of acid-deposition decisions and policy, we highly recommend that the work in developing and testing the EBA be continued.

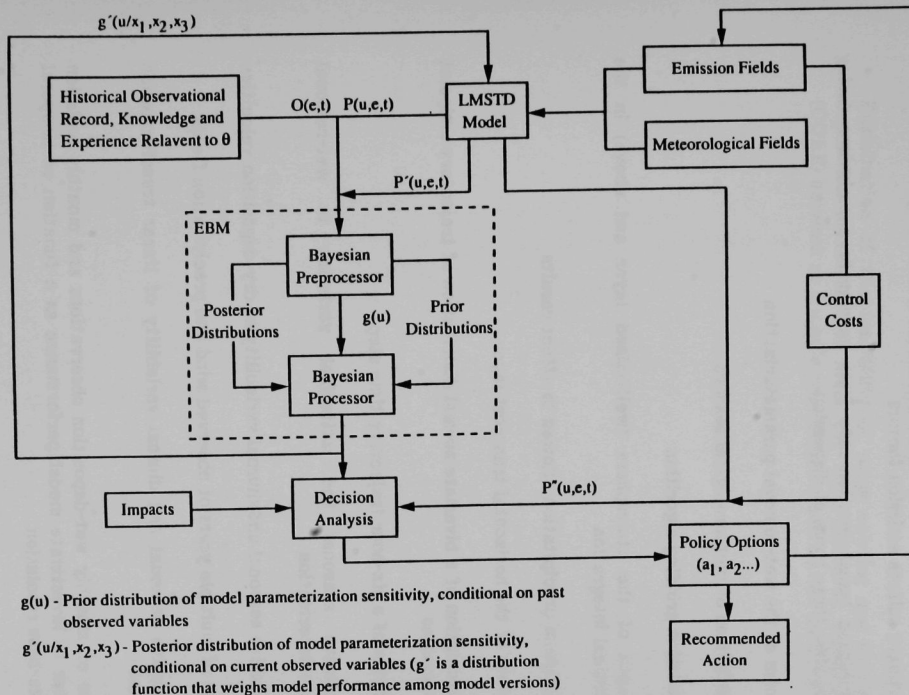
### 6.2.2 Estimation of Error in Data Bases

As mentioned throughout this report, unquantified errors in the model input data (i.e., emissions and meteorology) and errors in the model verification data (i.e., field-measured air concentrations and wet-deposition fluxes) make the evaluation of why a model fails or succeeds in reproducing observed spatial and temporal patterns extremely difficult. Attention must be given to work that will attempt to estimate this error and incorporate it into model evaluation performance measures. Most critical to improving the understanding and interpretation of model performance are the estimation of error in the model evaluation field data and the segregation (i.e., separation from model and model input data) of this error so it can be expressed as separate error elements in the computed model performance measures. One approach to the quantification and segregation of observational error is the use of a maximum likelihood and least-squares estimator as suggested by Jones (1979).

### 6.2.3 Model Sensitivity to Model and Data-Base Uncertainty Perturbations

If error bounds or limits in the model parameterizations and/or the processing and treatment of the model input data can be set, then, in principle, sensitivity to model predictions to local (one variable at a time) and global (all variables simultaneously) variations within the specified error bounds can be determined. This determination can be made by using methods such as global or local sensitivity analysis (with sets of ordinary or partial differential equations) and/or empirical orthogonal function analysis. In this study, we have only looked at model performance sensitivity in terms of variations in four internal model parameterizations ( $V_d$  for  $\text{SO}_2$  and  $\text{SO}_4^{2-}$ ,  $T_r$ , and WC). Several other model parameters or modeling assumptions (i.e., involving the use of higher-resolution model evaluation and/or input data and the way these data are processed or treated) not previously investigated can be varied. These could prove to be extremely useful in diagnosing the sources of error in model performance so that performance can be improved in a scientifically defensible manner. Some suggestions to consider are listed.

- Diurnal and seasonal patterns of vertical profiles of eddy diffusivity (an indicator of stability) and implied mixing heights



**FIGURE 6.1 Generalized Framework for Bayesian Uncertainty Analysis**

- Rate of loss of pollutant to the free troposphere by precipitation systems
- Upper and lower limits to dry deposition velocities during the first three hours of dispersion
- Primary sulfate emission factors
- Increased transformation rate from near-surface emissions during the first three hours of dispersion
- Form of the wet-removal parameterization
- Definition of meteorological seasons
- Emission gridding algorithm
- Depth of the atmosphere (well-mixed layer and above) in the vertical integration
- Minimum precipitation allowed to affect results
- Depth of the horizontal transport layer
- Assumption of a bivariate normal distribution of trajectory endpoint ensembles
- Choice of a six-hour trajectory time step
- Relative seasonal and latitudinal variation in wet-removal parameterization
- Relative seasonal and diurnal variability of dry-deposition velocities
- Use of multiple years of observed wind and precipitation fields
- Relative seasonal and diurnal variability of linear transformation rates
- Use of monthly wet-deposition observations and monthly emission rates to investigate model performance as a function of temporal data-base resolution
- Use of the most recent version of the NAPAP source emission inventory
- Variation in the spatial (horizontal and vertical) distribution and magnitude of emission fields

- Use of alternative methods to generate air parcel trajectories (i.e., numerical weather prediction models)
- Use of multi-year stochastically generated wind and precipitation fields to generate *forecasted responses* in source receptor relationships (Small 1985, see App. N)
- Examination of the feasibility of using existing and future dry-deposition measurements from the Core Research Establishment (CORE) network as a supplemental model evaluation data base

## REFERENCES

- Ball, R.H., 1986, U.S. Department of Energy, Office of Environmental Analysis, personal communication to M.A. Lazaro, Argonne National Laboratory, Oct.
- Ball, R.H., 1987, U.S. Department of Energy, Office of Environmental Analysis, personal communication to M.A. Lazaro, Argonne National Laboratory, April.
- Bilonick, R.A., 1985, *The Space-Time Distribution of Sulfate Deposition in the Northeastern United States*, Atmospheric Environment, 19(11):1829-1845.
- Briggs, G.A., 1971, *Some Recent Analysis of Plume Rise Observations*, Proc. 2nd International Clean Air Congress, H.M. Englund and W.T. Baery, eds., Academy Press, New York City, pp. 1029-1032.
- Chan, W.H., et al., 1985, *An Evaluation of Sampler Types and Sampling Periods for Measurement of Wet and Dry Deposition*, Ontario Ministry of the Environment Report ARB-98-85-AQM, Toronto.
- Clark, T.L., et al., 1987a, *International Sulfur Deposition Model Evaluation*, U.S. Environmental Protection Agency Report EPA/600/3-87/008, May.
- Clark, T.L., et al., 1987b, *The International Sulfur Model Evaluation Study*, presented at 80th Annual Meeting of Air Pollution Control Assn., New York City, June.
- Cox, W.M., et al., 1985a, *Preliminary Conclusions from EPA's Model Evaluation Program*, presented at 78th Annual Meeting of Air Pollution Control Assn., Detroit, June.
- Cox, W.M., and J.A. Tikvart, 1985b, *Assessing the Performance of Air Quality Models*, presented at 15th International Meeting on Air Pollution and Its Applications, CCMS/NATO, St. Louis, April.
- Dana, M.T., 1980, *SO<sub>2</sub> Versus Wet Deposition in the Eastern United States*, J. Geophysical Research, 85(C8):4475-4480.
- Dana, M.T., and R.C. Easter, 1987, *Statistical Summary and Analyses of Event Precipitation Chemistry from the MAP3S Network, 1976-1983*, Atmospheric Environment, 21(1):113-128.
- Dayan, U., et al., 1985, *An Assessment of Precipitation Chemistry Measurements from the Global Trends Network and Its Predecessors (1972-1982)*, Environmental Research Laboratory Report ERL ARL-136, Silver Springs, Md.
- Delhomme, J.P., 1978, *Kriging in the Hydrosiences*, Advances in Water Resources, 1(5):251.



Demerjian, K.L., 1985, *Quantifying Uncertainty in Long-Range Transport Models — A Summary of the AMS Workshop on Sources and Evaluation of Uncertainty in Long-Range Transport Models*, Bulletin of the American Meteorological Society, 66(12):1533.

Dennis, R.L. and S.K. Seilkop, 1986, *The Use of Spatial Patterns and Their Uncertainty Estimates in the Model Evaluation Process*, presented at 5th Joint Conf. on Applications of Air Pollution Meteorology, American Meteorological Society and Air Pollution Control Assn., Chapel Hill, N.C., Nov.

DePena, R.G., et al., 1985, *Wet Deposition Monitoring — Effect of Sampling Period*, Atmospheric Environment, 19(1):151.

Efron, E., and G. Gong, 1983, *A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation*, The American Statistician, 37 (Feb.).

Elliassen, A., 1978, *The OECD Study of Long Range Transport of Air Pollutants: Long Range Transport Modeling*, Atmospheric Environment, 12:479-487.

EPRI, 1983, *The Sulfate Regional Experiment: Report Findings*, Vol. 1-3 Electric Power Research Institute Report EPRI EA-1901.

EPRI, 1982, *The Sulfate Regional Experiment, Documentation of SURE Sampling Sites*, Electric Power Research Institute Report EPRI EA-1902.

EPRI, 1979, *Implementation and Coordination of the Sulfate Regional Experiment (SURE) and Related Research Programs*, Electric Power Research Institute Report EPRI EA-1066.

Eynon, B.P., and P. Switzer, 1983, *The Variability of Rainfall Acidity*, Canadian J. of Statistics, 11(1):11-24.

Fay, J.A., D. Golomb, and S. Kumar, 1985, *Source Apportionment of Wet Sulfate Deposition in Eastern North America*, Atmospheric Environment, 19(11):1773-1782.

Finkelstein, P.L., 1983, *The Spatial Analysis of Acid Precipitation Data*, J. of Climate and Applied Meteorology, 23 (Jan.):52-62.

Fox, D.G., 1984, *Uncertainty In Air Quality Modeling — A Summary of the AMS Workshop on Quantifying and Communicating Model Uncertainty*, Bulletin of the American Meteorological Society, 65(1):27-35.

Gardner, R.H., W.G. Cave, and R.V. O'Neill, 1982, *Robust Analysis of Aggregation Error*, Ecology, 63(6):1771-1779.

Goodin, W.R., G.J. McRae, and J.H. Selfeld, 1979, *A Comparison of Interpolation Methods for Sparse Data: Application to Wind and Concentration Fields*, J. of Applied Meteorology, pp. 761-771.

- Hakkarinen, C., 1982, *Comparing Data from Six North American Precipitation Chemistry Networks*, Proc. Specialty Conf. on Acid Deposition, Air Pollution Control Assn., Detroit, Nov.
- Hales, J.M., and M.T. Dana, 1979, *Regional-Scale Deposition of Sulfur Dioxide by Precipitation Scavenging*, Atmospheric Environment, 13:1121-1132.
- Hanna, S.R., and D.W. Heinold, 1985, *Development and Application of a Simple Method for Evaluating Air Quality Models*, Environmental & Research Technology, Inc., Report, Concord, Mass., March.
- Hidy, G.M., and P.K. Mueller, 1981, *A Digest of the Results from the Sulfate Regional Experiment (SURE)*, Proc. 74th Annual Meeting of Air Pollution Control Assn., June.
- Irwin, J., and M. Smith, 1984, *Potentially Useful Additions to the Rural Model Performance Evaluation*, Bulletin of the American Meteorological Society, 65(6).
- Isaacs, G.A., 1983, *The Vertical Transport and Redistribution of Pollutants by Clouds*, Transactions Specialty Conf. on the Meteorology of Acid Deposition, Air Pollution Control Assn., Hartford, Conn., Oct.
- Jones, T.A., 1979, *Fitting Straight Lines When Both Variables Are Subject to Error. I. Maximum Likelihood and Least-Squares Estimation*, J. of the International Assn. for Mathematical Geology, 11(1):1-25.
- Kleindiest, T.E., et al., 1988, *Comparison of Techniques for Measurement of Ambient Levels of Hydrogen Peroxide*, Environmental Science and Technology, 22:53-61.
- Knudson, D.A., 1985, *An Inventory of Monthly Sulfur Dioxide Emissions for the Years 1975-1983*, Argonne National Laboratory Report ANL/EES-TM-277, April.
- Kuo, Y.-H., et al., 1984, *The Accuracy of Air Parcel Trajectories as Revealed by Observing System Simulation Experiments*, Proc. 4th Joint Conf. on Applications of Air Pollution Meteorology, American Meteorological Society and Air Pollution Control Assn., Portland, pp. 228-231.
- Lazaro, M.A., D.G. Streets, and T. Surles, 1986, *A Two-Phase Framework for Acid Deposition Science-Policy Decisions*, presented at 7th World Clean Air Congress and Exhibition of the International Union of Air Pollution Prevention Assns., Sydney, Australia, Aug.
- Lee, Y.-N., et al., 1986, *Kinetics of Hydrogen Peroxide-Sulfur (IV) Reaction in Rainwater Collected at a Northeastern U.S. Site*, J. of Geophysical Research, 91(D12):13,264, Nov. 20.
- Liu, S.C., and J.R. McAfee, 1984, *Radon 222 and Tropospheric Vertical Transport*, J. of Geophysical Research, 89(D5):729-7297.

MacCracken, M.C., 1979, *The Multistate Atmospheric Power Production Pollution Study*, MAP3S, U.S. Department of Energy Report DOE/EV-0040.

Marnicio, R.J., et al., 1985, *A Comprehensive Modeling Framework for Integrated Assessments of Acid Deposition*, presented at 78th Annual Meeting of Air Pollution Control Assn., Detroit, June.

Matheron, G., 1971, *The Theory of Regionalized Variables and Its Applications*, Fontainebleau, 5.

McLarin, D.H., 1974, *Drawing Contours from Arbitrary Data Points*, Computational J., 17:318-324.

McNaughton, D.J., 1980, *Time Series Comparisons of Regional Model Predictions with Sulfur Dioxide Observations from SURE Program*, presented at 73rd Annual Meeting of Air Pollution Control Assn., Montreal, June.

McNaughton, D.J., C.M. Berkowitz, and R.C. Williams, 1981, *A Diagnostic Analysis of a Long-Term Regional Air Pollution Transport Model*, J. of Applied Meteorology, 20:795-801.

Mueller, P.K., and J.G. Watson, 1981, *The SURE Measurements*, Proc. 74th Annual Meeting of Air Pollution Control Assn., June.

MOI, 1982, *U.S.-Canada Memorandum of Intent on Transboundary Air Pollution*, Atmospheric Sciences and Analysis, Work Group 2, Regional Modeling Subgroup Report No. 2F-M, Nov.

NAPAP, 1985, *The National Acid Precipitation Assessment Program Annual Report to the President and Congress*, Interagency Task Force on Acid Precipitation, Washington, D.C.

NAPAP, 1987a, *The National Acid Precipitation Assessment Program Interim Assessment — The Causes and Effects of Acidic Deposition*, Vol. II — Emissions and Control, Interagency Task Force on Acid Precipitation, Washington, D.C.

NAPAP, 1987b, *The National Acid Precipitation Assessment Program Interim Assessment — The Causes and Effects of Acidic Deposition*, Vol. III: Atmospheric Processes, Interagency Task Force on Acid Precipitation, Washington, D.C.

Oden, N., 1986, *Kriging and Its Relations to Least Squares*, Brookhaven National Laboratory Report BNL 35614.

Olsen, A.R., et al., 1987, *Pacific Northwest Laboratory, Unified Wet Deposition Data Summaries for North America: Data Summary Procedures and Results*, to be published in Atmospheric Environment.

Peden, M., and L. Skowron, 1978, *Ionic Stability of Precipitation Samples*, Atmospheric Environment, 12:2343-2349.

Rabitz, H., M. Kramer, and D. Dacol, 1983, *Sensitivity Analysis in Chemical Kinetics*, Annual Review of Physical Chemistry, 34:419-61.

Ripley, B.D., 1981, *Spatial Statistics*, Wiley, New York City.

Rodda, J.C., S.N. Smith, and I.C. Strangeways, 1985, *On More Realistic Measurements of Rainfall and Their Impact on Assessing Acid Deposition*, ETH/IAHS/WMO Workshop on the Correction of Precipitation Measurements, Zurich, April.

Rodda, J.C., 1986, *Precipitation Research*, Transactions of the American Geophysical Union, Earth and Ocean Sciences Forum, Jan.

Ruff, R.E., et al., 1985, *Evaluation of Three Regional Air Quality Models*, Atmospheric Environment, 19:1103-1115.

Schiermeier, F.A., and P.K. Misra, 1982, *Evaluation of Eight Linear Regional-Scale Sulfur Models by the Regional Modeling Subgroup of the U.S./Canadian Memorandum of Intent Work Group*, Transactions Specialty Conf. on the Meteorology of Acid Deposition, Air Pollution Control Assn., pp. 330-345.

Schwartz, S.E., 1984, Vol. 3: *Gas-Phase Reactions of S and N Oxides in Liquid Water Clouds, in SO<sub>2</sub>, NO, and NO<sub>2</sub> Oxidation Mechanisms: Atmospheric Considerations*, in Acid Precipitation Review, J.G. Calvert, ed., Butterworth Press, Stoneham, Mass., p. 173.

Sevruk, B., 1982, *Methods of Correction for Systematic Error in Point Precipitation Measurement for Operational Use*, World Meteorological Organization Operational Hydrology Report No. 21.

Seilkop, S.K., 1983, *A Comparison of Universal and Simple Kriging in the Interpolation of Values from Nonstationary Processes*, prepared for U.S. Environmental Protection Agency under Contract No. 306379NASA.

Shannon, J.D., 1985, *User's Guide for the Advanced Statistical Trajectory Regional Air Pollution (ASTRAP Model)*, U.S. Environmental Protection Agency Report EPA/600/8-85/0160.

Shannon, J.D., 1981, *A Model of Regional Long-Term Average Sulfur Atmospheric Pollution, Surface Removal, and Net Horizontal Flux*, Atmospheric Environment, 15:689.

Sisterson, D.L., and G.T. Tissue, 1979, *Comparison of the Acidity of Rain Samples from ACM and PNL Precipitation Collectors*, Argonne National Laboratory Report ANL-78-65, Part IV.

Small, M.J., 1985, Carnegie-Mellon University, personal communication to M.A. Lazaro, Argonne National Laboratory, May.

Stewart, D.A., R.E. Morris, and M.K. Liu, 1983, *Evaluation of Long-Term Regional Transport Models of Interest to the National Park Service*, Systems Application, Inc., Final Report SYSAPP 83/215, San Rafael, Calif.

Stoiber, R.E., S.N. Williams, and L.L. Marlinconico, 1980, *Mount St. Helens, Washington, 1980 Volcanic Eruption: Magnitic Gas Component During the First 16 Days*, Science, 208:1258-1259.

Tilden, J.W., and J.H. Seinfeld, 1982, *Sensitivity Analysis of a Mathematical Model for Photochemical Air Pollution*, Atmospheric Environment, 16(6):1357-1364.

Topol, L.E., and M. Lev-On, 1986, *Comparison of Weekly and Daily Wet Deposition Sampling Results*, Electric Power Research Institute Report.

UDDBC, 1985a, *A Unified Wet Deposition Data Base for Eastern North America: Data Screening, Calculation Procedures, and Result for Sulfates and Nitrates (1980)*, Canadian Federal-Provincial Research and Monitoring Co-ordinating Committee Final Report.

UDDBC, 1985b, *A Unified Wet Deposition Data Base for Eastern North America: Addendum with Results for Sulfates and Nitrates (1980-1983)*, Canadian Federal-Provincial Research and Monitoring Co-ordinating Committee Report.

Venkatram, A., and J. Pleim, 1985, *Analysis of Observations Relevant to Long-Range Transport and Deposition of Pollutants*, Atmospheric Environment, 19(4):659-667.

Venkatram, A., P.K. Karamchandani, and J. Scire, 1986, *Data Base Priorities Identified by Regional Modeling Studies*, presented at 79th Annual Meeting of Air Pollution Control Assn., Minneapolis, June.

Wagner, J.K., et al., 1986, *Development of the 1980 Emissions Inventory*, GCA Corp. Draft Report GCA-TR-86-G, Bedford, Mass., Oct.

Wesely, M.L., and J.D. Shannon, 1984, *Some Factors that Affect the Deposition Rates of Sulfur Dioxide and Similar Gases on Vegetation*, J. of the Air Pollution Control Assn., 27:1110-1116.

Wesely, M.L., et al., 1985, *Measurements and Parameterization of Particulate Sulfur Dry Deposition over Grass*, J. of Geophysical Research, 90:2131-2143.

Willmott, C.J., 1981, *On the Validation of Models*, Physical Geography, 2(2):184-194.

Willmott, C.J., 1982, *Some Comments on the Evaluation of Model Performance*, Bulletin of the American Meteorological Society, 63(11).

Willmott, C.J., 1984, *On the Evaluation of Model Performance in Physical Geography*, in *Spatial Statistics and Models*, D. Reidal Publishing Co., Boston.

Winkler, R.L., 1972, *An Introduction to Bayesian Inference and Decision*, Holt, Reinhart and Winston, New York City.





X

